



UNIVERSITAT DE  
BARCELONA

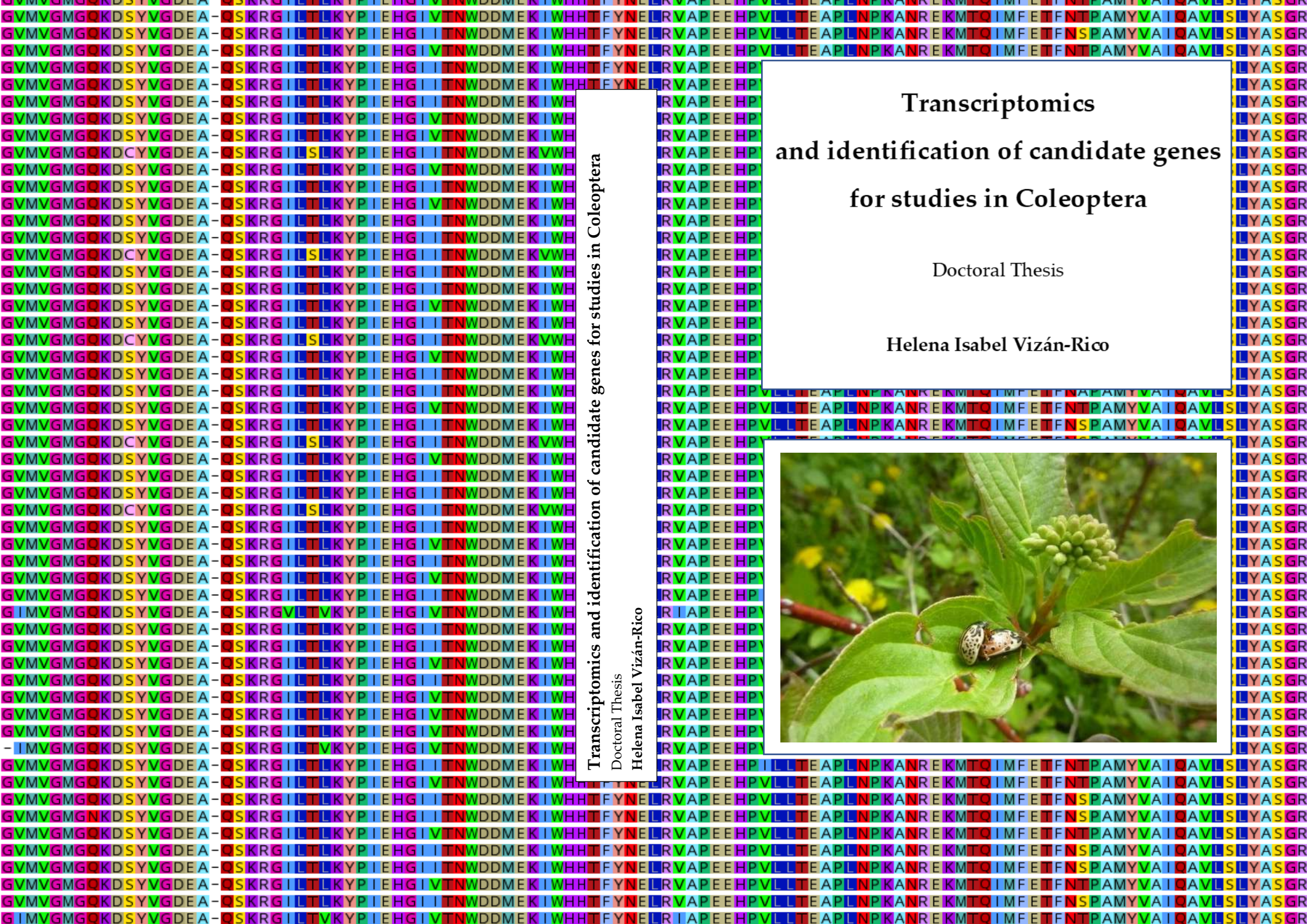
## Transcriptomics and identification of candidate genes for studies in Coleoptera

Helena Isabel Vizán-Rico

**ADVERTIMENT.** La consulta d'aquesta tesi queda condicionada a l'acceptació de les següents condicions d'ús: La difusió d'aquesta tesi per mitjà del servei TDX ([www.tdx.cat](http://www.tdx.cat)) i a través del Dipòsit Digital de la UB ([diposit.ub.edu](http://diposit.ub.edu)) ha estat autoritzada pels titulars dels drets de propietat intel·lectual únicament per a usos privats emmarcats en activitats d'investigació i docència. No s'autoritza la seva reproducció amb finalitats de lucre ni la seva difusió i posada a disposició des d'un lloc aliè al servei TDX ni al Dipòsit Digital de la UB. No s'autoritza la presentació del seu contingut en una finestra o marc aliè a TDX o al Dipòsit Digital de la UB (framing). Aquesta reserva de drets afecta tant al resum de presentació de la tesi com als seus continguts. En la utilització o cita de parts de la tesi és obligat indicar el nom de la persona autora.

**ADVERTENCIA.** La consulta de esta tesis queda condicionada a la aceptación de las siguientes condiciones de uso: La difusión de esta tesis por medio del servicio TDR ([www.tdx.cat](http://www.tdx.cat)) y a través del Repositorio Digital de la UB ([diposit.ub.edu](http://diposit.ub.edu)) ha sido autorizada por los titulares de los derechos de propiedad intelectual únicamente para usos privados enmarcados en actividades de investigación y docencia. No se autoriza su reproducción con finalidades de lucro ni su difusión y puesta a disposición desde un sitio ajeno al servicio TDR o al Repositorio Digital de la UB. No se autoriza la presentación de su contenido en una ventana o marco ajeno a TDR o al Repositorio Digital de la UB (framing). Esta reserva de derechos afecta tanto al resumen de presentación de la tesis como a sus contenidos. En la utilización o cita de partes de la tesis es obligado indicar el nombre de la persona autora.

**WARNING.** On having consulted this thesis you're accepting the following use conditions: Spreading this thesis by the TDX ([www.tdx.cat](http://www.tdx.cat)) service and by the UB Digital Repository ([diposit.ub.edu](http://diposit.ub.edu)) has been authorized by the titular of the intellectual property rights only for private uses placed in investigation and teaching activities. Reproduction with lucrative aims is not authorized nor its spreading and availability from a site foreign to the TDX service or to the UB Digital Repository. Introducing its content in a window or frame foreign to the TDX service or to the UB Digital Repository is not authorized (framing). Those rights affect to the presentation summary of the thesis as well as to its contents. In the using or citation of parts of the thesis it's obliged to indicate the name of the author.



Transcriptomics and identification of candidate genes for studies in Coleoptera

Doctoral Thesis

Helena Isabel Vizán-Rico

# Transcriptomics and identification of candidate genes for studies in Coleoptera

Doctoral Thesis

Helena Isabel Vizán-Rico





UNIVERSITAT DE  
BARCELONA



FACULTAD DE BIOLOGÍA  
DEPARTAMENTO DE GENÉTICA  
PROGRAMA DE DOCTORADO EN GENÉTICA

# Transcriptomics and identification of candidate genes for studies in Coleoptera

Transcriptómica  
e identificación de genes candidatos  
para estudios en Coleoptera

Memoria presentada por Helena Isabel Vizán-Rico  
para optar al título de Doctora por la Universidad de Barcelona

Trabajo realizado en el Institut de Biologia Evolutiva (CSIC-UPF)

**Helena Isabel Vizán-Rico**

Barcelona, 2021

Director

**Dr. Jesús Gómez-Zurita**  
Institut de Biologia  
Evolutiva (CSIC-UPF)

Tutor

**Dr. Pere Martínez Serra**  
Universitat de Barcelona  
(UB)

Doctoranda

**Helena Isabel Vizán-Rico**



*Gilgamesh, ¿a dónde vagas tú? La vida que persigues no hallarás. Cuando los dioses crearon la humanidad, la muerte para la humanidad apartaron, reteniendo la vida en las propias manos. Tú, Gilgamesh, llena tu vientre, goza de día y de noche. Cada día celebra una fiesta regocijada, ¡día y noche danza tú y juega! Procura que tus vestidos sean flamantes, tu cabeza lava; báñate en agua. Atiende al pequeño que toma tu mano, ¡Que tu esposa se deleite en tu seno! ¡Pues ésa es la tarea de la humanidad!*

Epopeya de Gilgamesh  
~2.500 a.C.

# 1. Prólogo

Como puede desprenderse de uno de los textos más antiguos conservados, de ~4.500 años de antigüedad, el ser humano ha estado siempre preocupado por la continuidad de la vida.

Desgraciadamente, y a pesar de la tecnología acumulada, el único remedio que puede hoy asemejarse, aunque de lejos, a ese fenómeno —y en un sentido amplio—, es la reproducción. Por supuesto, la reproducción no perpetuará la propia vida; pero sí, al menos, *La Vida*.

Tal vez el deseo inconsciente de perpetuidad (común) haya tenido algo que ver y haya empujado a la humanidad a escribir sobre infinitos temas, desde la ciencia hasta el arte y la filosofía, para conocer, escudriñar y tener a nuestro alcance todas las herramientas necesarias para comprenderlo todo.

El interés en la reproducción ha sido abordado desde múltiples campos de investigación y para múltiples finalidades. En este trabajo brindamos una aproximación al tema desde un punto de vista estrictamente genético y molecular, mediante el análisis de la evolución de genes responsables de funciones sexuales y restringido a un grupo de seres vivos: los insectos.

Para este objetivo, primero secuenciamos cinco transcriptomas de tejido reproductivo de macho (testículos y glándulas accesorias), analizamos su contenido y estructura, buscamos genes específicos y realizamos análisis filogenéticos y evolutivos.

## 2. Agradecimientos

El nacimiento de una tesis es un proceso complejo, multidisciplinar, que implica una gran cantidad y variedad de circunstancias convergentes, incluyendo personas, objetos, acciones e ideas, que se comportan como múltiples planos intersecando en múltiples líneas y puntos del espacio abstracto, generando formas caprichosas en dimensiones emergentes, como los planos que dibujan los cristales de una rosa del desierto.

Puesto que es sumamente difícil resumir todas/os las/os contribuyentes y los detalles de sus respectivas ayudas en un sencillo fragmento de papel, por favor, perdonadme con antelación; en compensación, haré lo que esté en mi mano para ser lo más cercana a la realidad.

Mi especial gratitud a Jesús Gómez-Zurita Frau por dirigir mi tesis doctoral, así como a Pere Martínez Serra y a Francesc Mestres por su ingente labor desde la Universitat de Barcelona (UB).

Mi gratitud especial a Christoph Mayer, por supervisar mi investigación en el Zoological Research Museum Alexander Koenig (ZFMK), sin cuya ayuda no habría podido llegar hasta aquí; a Bernhard Misof por darme la oportunidad de llevar a cabo una Estancia de Investigación en el ZFMK (y a mis tíos Dieter Höbel, Clara Vizán y Wunter Höbel por la ayuda de su intervención en alemán). Gracias a Christopher Mayer, Bernhard Misof y Oliver Niehuis por enseñarme a programar en Perl. Gracias a Hermes Escalona, Alejandro García Mondradón, Malte Petersen, y a todo el equipo del ZFMK por su cálida acogida.

Gracias a los colegas del todo el grupo de todo el *research lab* del Instituto de Biología Evolutiva CSIC-UPF (IBE): Tinguaro Montelongo, Gissela de la Cadena, Anna Papadopoulou, Nguyễn Thị Định, Ana Isabel, Josep Roca, así como a los colegas del IBE: Margarita Metallinou, Joan García-Porta, Guillem del Vallés, Helena Parra, Maria Rubio, Andrej Rudoy, David Sánchez, Amparo Hidalgo, Joao Maia, Luis Machado, Ana Trinidad, Marina Querejeta, José Luis Villanueva, Alfonso Balmori, Adrián Villastrigo, a todos los compañeros/as que ayudaron a la fundación del *IBE Student Group*, y a todos/as aquellos/as con los que mantení conversaciones científicas en el IBE o sus actividades.

Gracias a mis compañeros/as de máster de Biología Evolutiva Bárbara Simancas, Marta Cobo, Irene Cobo, Javier Santos, Marta, Natalia, Estefanía, Sergio, Mario,



Adhara Pardo, Adrián Páramo y Daniel Fernández; así como a Marta Cruz, Oriol Canals y demás compañeros de la UB. Gracias también a mis compañeros/as de estudios de Biología en la Universitat Autònoma de Barcelona (UAB): Noelia, Irene, Natalia y Arantxa.

Gracias a Javier Pérez-Tris por sus consejos en mi Trabajo de Fin de Máster en la Universidad Complutense de Madrid (UCM).

Gracias a Mariana Monteiro, del equipo de soporte de Blast2GO, por sus agudas explicaciones.

Gracias a Javier Tamames, por su ayuda en las búsquedas de *blast* locales al principio de mi tesis.

Gracias a todas/os las/os investigadoras/es con que me he encontrado en los diferentes congresos.

Gracias a todas/os mis profesoras/es, especialmente a los/as de Biología.

Gracias a las academias de arte Escola Joso, Escola Massana, Barcelona Academy of Arts (BAA) y Sant Lluc, por acompañar mi desarrollo y producción artística paralela en cómic, ilustración, y modelado, talla y escultura.

Gracias a mi familia (Jose M<sup>a</sup> Vizán Castro, M<sup>a</sup> Elena Rico Carranza, Blanca M<sup>a</sup> Vizán Rico y Nuria Sara Vizán Rico), por hacer que creciera mi mente hasta galaxias lejanas, sin límite ni dirección preestablecida; por regalarme interminables conversaciones alrededor la mesa, al despertar de las siestas vespertinas y durante emocionantes viajes y salidas; gracias por el sinfín de reuniones, entre amigos/as e invitados/as, en las que el rigor intelectual y el solaz la vida han danzado, en epicúreo equilibrio, en La Casa Abierta.

Gracias a Gerard, por su incondicional apoyo en todo momento mientras realicé esta tesis.



**Fig. 1.** Roca sedimentaria evaporítica comúnmente conocida como “Rosa del desierto”



### 3. Abstract

The molecular evolution of genes, joined to the differences between sexes, is a matter of interest poorly investigated across the class Insecta from a global perspective. The abundance of insect species, the high diversity of some of its orders, its ancient evolutionary origin, the current studies focused in a few model species, among other reasons, leads to a remarkable gap in studies in insect evolution and diversification, especially on beetles.

In this doctoral thesis, we sequenced by the first time five testis transcriptomes of four species of chrysomelids (*Calligrapha confluens*, *Calligrapha aff. floridana*, *Calligrapha multipunctata* and two specimens of *Calligrapha philadelphia* from two distant localities (PA and QC)) available from project CGL2011-23820, supporting this thesis. After the *de novo* assembly of the transcript contigs, functionally annotated 32,8-44,6% of them using the more suitable platform for *de novo* annotation of non-model species: Blas2GO.

We subsequently explored the Gene Ontology Consortium, searching for candidates for gene-finding. We selected 44 sperm individualization genes (GO:0007291) and explored their sex-biased condition in the insect model species *Drosophila melanogaster* through Flybase. Seven exhibited male-biased expression (*blanks*, *Cyt-c-d*, *gudu*, *hmw*, *klhl10*, *nsr* and *Prosalpha6T*) while one exhibited female-biased (*scat*).

Afterwards, we searched 20 beetle transcriptomes (19 of them from the 1KITE insect database), retrieving the putative orthologs of the 44 sperm individualisation genes. First, we confirmed the orthology of *CG9313*, *Tektin-A* and *tomboy40* in *Calligrapha* transcriptomes. Subsequently, we inferred gene trees of 41 sperm individualisation genes (three of them excluded) across the class Insecta (up to 119 species). We identified duplications in some lineages of the 41 gene trees, as well as putative events of gene loss.

Moreover, we estimated the evolutionary rates for each of the 41 sperm individualisation genes. For amino acid sequences of insect species, it resulted, in average,  $0.00239 \pm 0.003012$  subs./l./Ma (ranging from 0.000237 in *orb2* to 0.009667 in *hmw*). Regarding nucleotide sequences of beetle species, it resulted, in average,  $0.00452 \pm 0.002083$  subs./l./Ma (ranging from 0.00208 subs./l./Ma in *nes* to 0.01190 subs./l./Ma in *Cul3*; the time constraint for Coleoptera: 277.4-315.2

Ma).

We also searched for patterns of evolution of the 41 sperm individualisation genes. We found faster rates of evolution in proteins of sex-biased and clock-constrained genes (except for *hmv*, which is far from a molecular clock). However, we did not find faster evolution rates for the nucleotide sequences of the same genes for the conditions of sex-biased, presence of duplications or their position in an interaction network.

We also analysed a possible effect of gene interaction on the evolutionary rates of genes. Although we did not find significant differences in rate evolution between groups of interacting genes, we found significance in the edge *Dronc-shi* separating groups, as well as in the edges *Dronc-Dredd* and *Chc-Past1*.

In summary, through this doctoral thesis we have enlarged the genetic pool of insect species available on public databases of genes, enriching the representativeness of Coleoptera species; we have provided important information about the functionality of new sequenced testes-expressed genes; we have reported bioinformatic tools for gene-finding (by the use, development or improvement of them); we have contributed a considerable amount of gene trees reflecting the molecular evolution of sperm individualisation genes in insects; and, finally, we have investigated the patterns of evolution associated to different gene traits.

However, we still find an undoubtable missing data in our investigations; the new sequencing of insect species is slow so far, although it has increased in the last recent years. Also, studies on gene expression are currently lacking out of model species, either of sex, tissue or stage of development although a great effort is being done creating new databases of insect expression data. Altogether, further analyses are required to improve the picture of evolution of insect sex-specific expressed genes.

## 4. Índice

	Pág.
<b>1. Prólogo</b>	2
<b>2. Agradecimientos</b>	3
<b>3. <i>Abstract</i></b>	5
<b>4. Índice</b>	7
<b>5. Lista de tablas y figuras</b>	9
<b>6. Lista de abreviaturas</b>	6
<b>7. Introducción</b>	13
7.1. <i>Gene-finding: La búsqueda de genes</i>	14
7.2. <i>Las genotecas de EST tradicionales y sus limitaciones</i>	18
7.3. <i>El RNA-Seq</i>	19
7.4. <i>La expresión génica</i>	24
7.5. <i>Genes sex-biased en Insecta</i>	27
<b>8. Objetivos</b>	29
8.1. <i>Objetivo general</i>	30
8.2. <i>Objetivos específicos</i>	31
<b>9. Informe del factor de impacto de las publicaciones</b>	33
<b>10. Publicaciones</b>	35
10.1. <i>Publicación de primer autor nº1</i>	36
10.1. <i>Publicación de primer autor nº2</i>	50
<b>11. Discusión global</b>	75
11.1. <i>Obtención del material fuente: secuenciación de cinco transcriptomas de cuatro especies no modelo del género Calligrapha</i>	77

<b>11.2. Anotación funcional de novo de cinco transcriptomas del género <i>Calligrapha</i></b>	<b>79</b>
11.2.1. Comparación intra- e intertranscriptómica de la redundancia funcional	81
<b>11.3. Gene-finding en transcriptomas de especies de <i>Insecta</i></b>	<b>83</b>
11.3.1. Elección de genes diana	83
11.3.1.1. Primera aproximación	83
11.3.1.2. Segunda aproximación: representantes del Gene Ontology GO:0007291 (sperm individualization)	83
11.3.2. <i>Gene-finding</i> en los cinco transcriptomas de <i>Calligrapha</i>	85
11.3.3. <i>Gene-finding</i> en la base de datos de transcriptomas de insectos 1KITE	92
<b>11.4. Análisis filogenético</b>	<b>94</b>
11.4.1. Análisis filogenético de los genes de función masculina <i>CG9313</i> , <i>Tektin-A</i> y <i>tomboy40</i> en <i>Insecta</i>	94
11.4.2. Análisis filogenético de genes de individualización de espermatozoides (GO:0007291) en <i>Insecta</i> y <i>Coleoptera</i>	94
<b>11.5. Evolución molecular de genes de individualización de espermatozoides (GO:0007291) en <i>Insecta</i> y <i>Coleoptera</i></b>	<b>96</b>
<b>11.6. Redes de interacción génica de genes de individualización de espermatozoides (GO:0007291) en <i>Insecta</i> y <i>Coleoptera</i></b>	<b>99</b>
<b>12. Conclusiones</b>	<b>102</b>
<b>13. Bibliografía</b>	<b>106</b>
<b>14. Anexo</b>	<b>115</b>

## 5. Lista de tablas y figuras

	Pág.
<b>1. Prólogo</b>	
Fig. 1. Roca sedimentaria evaporítica comúnmente conocida como “Rosa del desierto”	4
<b>2. Agradecimientos</b>	
<b>3. Abstract</b>	
<b>4. Índice</b>	
<b>5. Lista de tablas y figuras</b>	
<b>6. Lista de abreviaturas</b>	
Tabla 1. Lista de abreviaturas y términos en inglés utilizados a lo largo del texto.	11
<b>7. Introducción</b>	
Tabla 2. Comparación de las características de los secuenciadores de segunda y tercera generación	17
Tabla 3. Resumen de ensambladores disponibles: tecnologías que los soportan, enlace de acceso y comentarios (Niranjan & Pop, 2013)	20
Fig. 2. El ensamblaje del genoma a partir de fragmentos cortos de ADN secuenciados (Baker, 2012)	21
Fig. 3. Comparación de ensambladores de novo para el éxito en la reconstrucción de genes completos (Zhao et al. 2011)	22
<b>8. Objetivos</b>	
<b>9. Informe del factor de impacto</b>	
<b>10. Publicaciones</b>	
<b>11. Discusión global</b>	

**Fig. 4.** Relaciones entre términos de Gene Ontology (GOs) en torno al término GO *spermatogeneis* (GO:0007283) 82

**Tabla 4.** Resumen de genes en los que se ha detectado respectivamente selección positiva o selección purificadora 97

**Fig. 5.** Perfil de expresión transcriptómica por tejidos del gen *gudu* en *Drosophila*, según los informes acumulados en el proyecto modENCODE de la base de datos de Flybase (Attril et al. 2016) 99

## 12. Conclusiones

**Fig. 6.** Ley cuadrático-cúbica, enunciada por Galileo Galilei en Dos nuevas ciencias: Discorsi e Dimostrazioni Matematiche, intorno a due nuove scienze (1638) 101

## 13. Referencias

## 14. Anexo

## 6. Lista de abreviaturas

A lo largo del presente texto aparecen numerosas abreviaturas y términos en inglés, para aclaración de los cuales se expone la siguiente tabla de sinónimos:

**Tabla 1.** Lista de abreviaturas y términos en inglés utilizados a lo largo del texto.

Abreviatura en inglés	Término en inglés	Término en castellano
	<i>Assembly</i>	Ensamblaje
	<i>Assembled</i>	Ensamblado(s)
	<i>Arrhenotokous haplodiploidy</i>	Haplodiploidía arrenotóquica
	<i>Arrhenotoky</i>	Arrenotoquía
	<i>Array(s)</i>	Matriz(es) (referido a matriz de datos)
	<i>Biased</i>	Sesgado/a
BP	<i>Biological Process</i>	Proceso Biológico (referido al dominio de <i>Gene Ontology Consortium</i> )
bp	<i>Base pair</i>	Par de bases
CC	<i>Cellular Component</i>	Componente Celular (referido al dominio de <i>Gene Ontology Consortium</i> )
	<i>Child terms</i>	Términos hijos (referido a los términos de ontología génica, GOs)
	<i>Enrichment</i>	Enriquecimiento (por lo general, referido a la proporción de genes expresados en un tejido versus el cuerpo entero del espécimen)
	<i>Expression array</i>	Matriz (de datos) de expresión
EST(s)	<i>Expressed Sequence Tag(s)</i>	Marcador(es) de secuencia expresada
	<i>Deep-sequencing technologies</i>	Tecnologías de secuenciación de alto rendimiento
E-value	<i>E-value</i>	Valor E (referido a las búsquedas de <i>blast</i> )
DNA	<i>Desoxyribonucleic Acid</i>	Ácido desoxirribonucleico (ADN)
	<i>Gene-expression</i>	Expresión génica
	<i>Gene finding/Gene-finding</i>	Búsqueda de genes
	<i>Female-biased genes</i>	Genes de expresión sesgada en hembras
w	<i>Fitness (w)</i>	Eficacia biológica (w)
GO	<i>Gene Ontology</i>	Ontología génica (OG)
	<i>Gene Ontology Consortium</i>	Proyecto de Ontología Génica (OG)
	<i>Haplodiploidy</i>	Haplodiploidía
	<i>High-throughput sequencing technologies</i>	Tecnologías de secuenciación de alto rendimiento
	<i>Male-biased genes</i>	Genes de expresión sesgada en machos



	<i>Male-function genes</i>	Genes de función masculina
	<i>Mapping</i>	Mapeo
mRNA	<i>Messenger Ribonucleic Acid</i>	Ácido ribonucleico mensajero (ARNm)
	<i>Missing data</i>	Datos que faltan
MF	<i>Molecular Function</i>	Función Molecular (referido al dominio de <i>Gene Ontology Consortium</i> )
MYA	<i>Millions Years Ago</i>	Millones de años (Ma)
NGS	<i>Next Generation Sequencing</i>	Secuenciación de Nueva Generación
NGS technologies	<i>Next Generation Sequencing technologies</i>	Tecnologías de Secuenciación de Nueva Generación
nt	<i>Nucleotide(s)</i>	Nucleótido(s)
	<i>Parthenogenesis</i>	Partenogénesis
	<i>Pipeline</i>	Fuente de información
	<i>Pool (genetic pool)</i>	Acervo (acervo genético)
PCR	<i>Polymerase Chain Reaction</i>	Reacción en cadena de la polimerasa
PGE	<i>Paternal genome elimination</i>	Eliminación de Genoma Paterno
	<i>Present call</i>	Señal presente
	<i>Primer(s)</i>	Cebador(es)
qPCR Q-PCR	<i>Quantitative Polymerase Chain Reaction</i>	Reacción en cadena de la polimerasa cuantitativo
	<i>Raw read(s)</i>	Lectura(s) cruda(s) (referido a la lectura de secuencias obtenidas mediante NGS)
	<i>Read(s)</i>	Lectura(s) (referido a la lectura de secuencias obtenidas mediante NGS)
	<i>Regular expression</i>	Expresión regular (uso en programación)
RNA	<i>Ribonucleic Acid</i>	Ácido ribonucleico (ARN)
RNAlater	<i>Ribonucleic Acid later</i>	Conservante para ARN (Qiagen)
RNA-Seq	<i>RNA-Sequencing</i>	Secuenciación de ARN
RT-PCR	<i>Real Time Polymerase Chain Reaction</i>	Reacción en cadena de la polimerasa a tiempo real
	<i>Reporter(s)</i>	Informe, estudio
S.D.	<i>Standard Deviation</i>	Desviación típica o desviación estándar ( $\sigma$ , s)
	<i>Sex-bias</i>	Sesgo entre machos y hembras
	<i>Sex-biased</i>	Sesgado/a/os/as entre machos y hembras
	<i>Sex-biased gene(s)</i>	Gene(s) de expresión sesgada entre machos y hembras
	<i>Short read(s)</i>	Secuencia(s) corta(s) (referido a secuencias procedentes de secuenciación)
	<i>Transcripts</i>	Tránscritos

## **7. Introducción**

## 7.1. Gene-finding: *La búsqueda de genes*

La obsesión científica por el completo discernimiento del *sustrato de la herencia* no ha dejado de crecer desde sus comicios, desde que se hizo figura de él —aunque fuera de forma abstracta— en las extensas páginas sobre las variaciones individuales y los problemas para delimitar especies basados la variación fenotípica de Charles Darwin en *El origen de las especies* (1859), hasta su real secuenciación en 1965 con Robert Holley et al. —la alanine-tRNA de *Saccharomyces cerevisiae*— seguido de una productiva década con la secuenciación de Fiers en 1972 —510 bp del gen de la proteína de la cubierta del virus de ARN bacteriófago MS2 (Min Jou et al. 1972)—, con Gilbert y Maxam en 1973, y con Sanger y colaboradores inmediatamente después, en 1975, cuyo método definitivamente se popularizaría y marcaría la era de la secuenciación.

Sin embargo, dado que el progreso científico no se detiene, la mejora de los métodos de secuenciación tampoco; y, así, la era de la secuenciación ha sufrido tres grandes cambios, mejoras o transformaciones:

- (i) La primera generación de secuenciación, caracterizada por el uso del método Sanger y el de Maxam & Gilbert, marcada por la optimización en 1977 del método Sanger (utilizando un procedimiento de terminación de cadena conocido como método de los didesoxinucleótidos, en el que diversos didesoxirribonucleótidos trifosfato (ddATP, ddGTP, ddCTP y ddTTP), al ser incorporados por el enzima DNA polimerasa, impiden la adición de más desoxirribonucleótidos trifosfato (dATP, dGTP, dCTP y dTTP)) o el de Gilbert (basado en el método de fragmentación química).
- (ii) La segunda generación de secuenciación, caracterizada por el uso de secuenciadores de alto rendimiento o *high-throughput*, capaces de secuenciar cientos de miles de moléculas en paralelo gracias a la inmovilización de las reacciones sobre una superficie sólida.
- (iii) La tercera generación, *single molecule real time sequencing*, basada en la secuenciación a tiempo real de miles de millones de pequeñas moléculas de ADN adheridas a una superficie sólida, con la

fundamental diferencia de que, en este caso, las reacciones contienen moléculas únicas.

El método Sanger, una vez optimizado (recordemos que originalmente requería diferentes pasos con tediosa impliación manual, tales como electroforesis en gel de poliacrilamida, transferencia Southern con marcaje por radioactividad, revelado inmunológico de bandas, lectura manual de placas de secuenciación, etc.), acabó por dominar la era de la secuenciación de primera generación, relegando a un segundo plano el método de Maxam & Gilbert basado en el método de fragmentación química. El resultado era un cromatograma del que se deducía la secuencia mediante el ordenador. Para obtenerlo utilizaba didesoxinucleótidos marcados con fluorescencia que se analizaban en una electroforesis capilar y producían un cromatograma o electroferograma, a partir del cual se deducía la secuencia en el ordenador.

Así, los secuenciadores automáticos, *ABI Prism (Applied Biosystems)* o de *CEQ-serie (Beckman Coulter)*, consiguen secuenciar simultáneamente hasta 96 muestras de ADN de 500-1000 bases de longitud en pocas horas.

La pirosecuenciación de la segunda generación, en cambio, utiliza una técnica diferente no fluorescente: se mide la liberación de pirofosfato en una reacción de polimerización mediante una serie de reacciones enzimáticas acopladas que liberan luz cada vez que se incorpora un nucleótido. El resultado es una imagen que, tras ser analizada e interpretada en el ordenador (a través de flujogramas), resuelve la secuencia nucleotídica.

El primer modelo comercializado de esta tecnología fue el GS20 (de la empresa *454 Life Sciences*, posteriormente absorbida por *Roche*), el cual podía secuenciar hasta 20 millones de bases en 4 h.

Pero otras dos compañías, *Solexa* y *SOLiD* (de *Applied Biosystems*) desarrollaron paralelamente otros dos métodos no basados en la pirosecuenciación:

- a) *Solexa* (Metzker, 2010) utiliza un método de la polimerización del ADN particular. Cada nucleótido que se incorpora a la cadena está marcado mediante fluorescencia y protegido en la cadena naciente, de manera que impide que ésta siga creciendo. Al detectarse la señal fluorescente, se elimina la protección del nucleótido, permitiendo la incorporación de otro nucleótido marcado. El ciclo se repite sucesivamente.

- b) La tecnología de SOLiD (*Applied Biosystems*) es similar, ya que también detecta la señal fluorescente tras cada adhesión, pero con la diferencia de que utiliza la ligación de octámeros marcados de secuencia conocida a la cadena de ADN.

Ambas tecnologías aventajan a la pirosecuenciación en que resuelven de forma fiable las regiones homopoliméricas del ADN. Otra ventaja importante reside en que son capaces de secuenciar el ADN sin el paso previo de la clonación, por lo que no es necesaria la creación de genotecas, la replicación en *Escherichia coli*, la purificación ni la conservación a -80°C. Por último, poseen asimismo la ventaja de un bajo coste económico.

Sin embargo, poseen un grave inconveniente, y es que no son capaces de generar lecturas de muchas pares de bases (originalmente 75bp, aunque han ido optimizándose y cada vez se consiguen obtener lecturas superiores (~300bp)), por lo que no son útiles si nos interesa la secuenciación *de novo*. Son adecuados para resecuenciar genomas ya conocidos o analizar la expresión de los *transcripts*.

De entre los secuenciadores de tercera generación (*single molecule real time sequencing*), que secuencian cada molécula de ADN individualmente, cabe destacar:

- a) *Helicos BioSciences*, el primero que se desarrolló, que consigue lecturas fiables de entre 25-45 bases. Por la cortedad de las secuencias es recomendable, de nuevo, para resecuenciaciones, pero no para secuenciación *de novo*.
- b) *PacificBiosciences*, que obtendría secuencias de hasta 1000 nt mediante el anclaje a la superficie sólida de la ADN polimerasa.
- c) *ZS Genetics*, que utiliza microscopía electrónica para leer la secuencia de ADN directamente sobre una imagen electrónica, previa replicación de la hebra molde de ADN mediante bases yodadas, bromadas o triclorometiladas.

**Tabla 2.** Comparación de las características de los secuenciadores de segunda y tercera generación

	Equipo	Compañía	Método de secuenciación	DNA molde	Longitud lecturas (pb)	Tiempo carrera (h)	nt/carrera (Gb)
<b>2ª Generación de secuenciadores</b>	GS-FLX (454)	Roche	Polimerasa (pirosecuenciación)	PCR Emulsión	250-400	10	0,4
	SOLEXA	Illumina	Polimerasa (terminadores reversibles)	PCR Puente	35-75	48	18
<b>3ª Generación de secuenciadores</b>	ABI SOLID	Applied Biosystems	Ligasa (octámeros con código de dos bases)	Emulsión	25-75	168	30
	Helicos tSMS	Helicos BioSciences	Polimerasa	Molécula única	25-45	192	30
	Pacific Biosciences	Pacific Biosciences	Polimerasa	Molécula única	1000	NA	NA
	ZX Genetics	ZX Genetics	Microscopia electrónica	Molécula única	NA	NA	NA

Durante este período, y mientras un número creciente de secuencias puebla las bases de datos alrededor del mundo, nuevas cuestiones aparecen en el campo de la investigación, que fuerzan el desarrollo de nuevos métodos; los cuales, siguiendo un esquema de retroalimentación positiva, generan nuevos datos sobremanera, los cuales se almacenan en bases de nueva creación.

Se alimenta de este modo una rueda que bebe de la información pasada y de la cuestión nueva, que avanza abandonando las técnicas viejas y aquellas concepciones erróneas, pero que nunca se desinfla, sino que siempre crece.

Sin embargo, y paradójicamente, la cuestión de la masificación de la información provoca *per se* que muchas de las nuevas preguntas no puedan abordarse de manera directa, y que se requieran herramientas para el filtrado de datos, ya sea con el objetivo de focalizar el trabajo en una fracción del total de la información disponible, o bien, con el de abordar su totalidad de forma completa y transversal.

La estrategia *gene-finding*, que consiste en la identificación de regiones del genoma, previa secuenciación, que codifican para genes funcionales podría, debería o intentaría en tal modo, discernir parte de estas cuestiones.

Mucho se ha avanzado desde los primeros experimentos de *gene-finding* realizados en células vivas u organismos, pasando por la obtención de *Expressed Sequence Tags* (ESTs) hasta las recientes tecnologías de *Next Generation Sequencing* (NGS), especialmente el RNA-Seq y el CHIP-Seq.

## 7.2. Las genotecas de EST tradicionales y sus limitaciones

Un marcador de secuencia expresada o EST (del inglés, *Expressed Sequence Tags*) es una secuencia corta (generalmente, de menos de 1000bp) de cDNA, procedente de un solo paso de lectura del mRNA (NCBI; available from: <https://www.ncbi.nlm.nih.gov/genbank/dbest/>).

Suelen secuenciarse de forma masiva, obteniéndose un cuadro del conjunto de genes expresados en un individuo, en un tejido concreto o en una etapa específica de su desarrollo.

A estos conjuntos de secuencias expresadas las denominamos bibliotecas de cDNA, y ha sido una de las técnicas más utilizadas durante las últimas décadas para la exploración de la secuencia genómica. Entre sus aplicaciones destacan el campo de la filogenética, la identificación de perfiles transcriptómicos y la proteómica (Parkinson & Blaxter, 2009).

La generación de secuencias ESTs se generalizó enormemente en las pasadas décadas, llenando las bases de datos de genes como la de GenBank del NCBI (Geer et al. 2010), la cual empezó a crecer exponencialmente. Tanto fue así, que en 1992 el NCBI incentivó la creación de la base de datos para almacenar la información específica de secuencias ESTs.

En 1993, Genbank creó dbEST, una sección específica para los EST de un cierto número de especies (Boguski, 1993) y, desde entonces, no ha dejado de crecer.

A pesar de todo, diez años después, la tasa de crecimiento era todavía demasiado lenta para alcanzar estudios genéticos a larga escala taxonómica; e importantemente, estaba significativamente sesgada en cuanto a especies, legando a las especies no modelo a una considerable infrarrepresentación, como es el caso de las pertenecientes al orden Coleoptera.

De esta manera, en 2002 había solamente 6.147 secuencias en el NCBI pertenecientes a coleópteros y, restringiéndose a genes de coleópteros no mitocondriales o ribosomales, el número se reducía a 200 (Theodorides et al. 2002).

La llegada del NGS ha vuelto a forzar un cambio en las bases de datos de GenBank, el cual ha creado un canal exclusivo para el almacenaje de ESTs procedentes de secuenciación paralela, conocido como *Sequence Read Archive* o SRA (NCBI; available from: <https://trace.ncbi.nlm.nih.gov/Traces/sra/sra.cgi>).



### 7.3. El RNA-Seq

El RNA-Seq, la reciente técnica desarrollada para el análisis de transcriptomas mediante *Next generation sequencing technologies* (Wang et al. 2009), se ha convertido en una de las técnicas líder actualmente utilizadas para el *gene-finding*. El almacenaje de genes expresados en librerías de secuencias de RNA-Seq constituye una buena base como material fuente de múltiples estudios.

Las librerías obtenidas por RNA-Seq consisten en colecciones de secuencias de DNA complementario (cDNA) obtenido de una mayoría —idealmente, todos— los fragmentos de mRNA presentes en una célula, tejido, organismo o muestra ambiental, secuenciados de forma masiva: por lo general, se obtienen del orden de ~100.000 secuencias limpias por muestra.

El proceso de obtención ha de ser rápido y meticuloso, ya que el mRNA se degrada rápidamente, debido a la acción de las RNasas, presentes prácticamente en todos los organismos y también libres en el ambiente. Por tanto, las muestras deben ser obtenidas *in situ* y conservadas en RNAlater (Invitrogen) para ser posteriormente secuenciadas. A pesar de lo anterior, ciertos intentos por obtener mRNA de muestras de insectos conservadas en alcohol durante años han sido exitosos (Bernhard Misof, comunicación personal).

La metodología del RNA-Seq procedente NGS ha relegado a un segundo plano los antiguos métodos de secuenciación Sanger, ya que ofrece enormes ventajas respecto a este. Marguerat y Bähler (2010), dos años después de las primeras aplicaciones del RNA-Seq, ya hablaban de sus importantes contribuciones en el discernimiento de la expresión del genoma y su regulación.

A modo de ejemplo, Azevedo et al. (2012) reporta 3.068 nuevos ESTs de la vesícula seminal de *Lutzomyia longipalpis*, de los cuales solo 2.678 serían de buena calidad y solo 1.391 serían únicos (no redundantes); en cambio, con la técnica de RNA-Seq, se obtienen del orden de 100.000 secuencias por muestra, las cuales, aunque no corresponden a secuencias únicas, suponen un número incomparablemente superior que no admite parangón.

A pesar de todo, los métodos de RNA-Seq también presentan inconvenientes. El más inmediato tal vez sea que la tecnología requiere un tratado posterior de las secuencias cortas obtenidas de la secuenciación: el ensamblado (o *assembly*). El ensamblado consiste en la unión de las secuencias crudas obtenidas de la secuenciación paralela (o *raw reads*) con el objetivo de reconstruir la secuencia

nucleotídica original y obtener así una librería de genes expresados por el espécimen (o tejido o célula de él). Para ello existen diversas herramientas bioinformáticas (Tabla 2).

**Tabla 3.** Resumen de ensambladores disponibles: tecnologías que los soportan, enlace de acceso y comentarios (Niranjan & Pop, 2013)

Assemblers	Technology	Availability	Notes
<b>Genome assemblers</b>			
<i>ALLPATHS-LG</i>	Illumina, Pacific Biosciences	<a href="ftp://ftp.broadinstitute.org/pub/crd/ALLPATHS/Release-LG">ftp://ftp.broadinstitute.org/pub/crd/ALLPATHS/Release-LG</a>	Requires a specific sequencing recipe
<i>SOAPdenovo Illumina</i>	Illumina	<a href="http://soap.genomics.org.cn/soapdenovo.html">http://soap.genomics.org.cn/soapdenovo.html</a>	Also used for transcriptome and metagenome assembly
<i>Velvet</i>	Illumina, SOLiD, 454, Sanger	<a href="http://www.ebi.ac.uk/~zerbino/velvet">http://www.ebi.ac.uk/~zerbino/velvet</a>	May have substantial memory requirements for large genomes
<i>ABYSS</i>	Illumina, SOLiD, 454, Sanger	<a href="http://www.bcgsc.ca/platform/bioinfo/software/abyss">http://www.bcgsc.ca/platform/bioinfo/software/abyss</a>	Also used for transcriptome assembly
<b>Metagenome assemblers</b>			
<i>Genovo</i>	454	<a href="http://cs.stanford.edu/group/">http://cs.stanford.edu/group/</a>	Uses a probabilistic model for genovo
<i>MetaVelvet</i>	Illumina, SOLiD, 454, Sanger	<a href="http://metavelvet.dna.bio.keio.ac.jp">http://metavelvet.dna.bio.keio.ac.jp</a>	Based on Velvet
<i>Meta-IDBA</i>	Illumina	<a href="http://i.cs.hku.hk/~alse/hkubrg/projects/metaidba">http://i.cs.hku.hk/~alse/hkubrg/projects/metaidba</a>	Based on IDBA
<b>Transcriptome assemblers</b>			
<i>Trinity</i>	Illumina, 454	<a href="http://trinityrnaseq.sourceforge.net">http://trinityrnaseq.sourceforge.net</a>	Tailored to reconstruct full-length transcripts; may require substantial computational time
<i>Oases</i>	Illumina, SOLiD, 454, Sanger	<a href="http://www.ebi.ac.uk/~zerbino/oases">http://www.ebi.ac.uk/~zerbino/oases</a>	Based on Velvet
<b>Single-cell assemblers</b>			
<i>SPAdes</i>	Illumina	<a href="http://bioinf.spbau.ru/en/spades">http://bioinf.spbau.ru/en/spades</a>	
<i>IDBA-UD</i>	Illumina	<a href="http://i.cs.hku.hk/~alse/hkubrg/projects/idba_ud">http://i.cs.hku.hk/~alse/hkubrg/projects/idba_ud</a>	Based on IDBA

Los programas de ensamblaje trabajan con secuencias de dos tipos como datos de entrada, conocidos como *single reads* y *paired reads*, que provienen de los *raw reads* tras el proceso de *cleaning*:

- i) Las *single reads* son simplemente fragmentos cortos secuenciados en sí mismos. Se combinan posteriormente (gracias a regiones que se superponen) en secuencias más largas que se denominan *contigs*.
- ii) Las *paired reads* tienen aproximadamente la misma longitud que las anteriores; pero, a diferencia de ellas, provienen de algún extremo del fragmento de ADN que era demasiado largo para ser secuenciado directamente. La distancia puede variar de 200bp a varias decenas de kilobases. El hecho de que provengan de una secuencia dividida puede ayudar en el proceso de unión de estas en andamios o *scaffolds*, es decir, conjuntos ordenados de *contigs* con *gaps* entre ellos. Las *paired reads* también ayudan a indicar el tamaño de las regiones repetitivas y la distancia entre *contigs* (Baker, 2012).

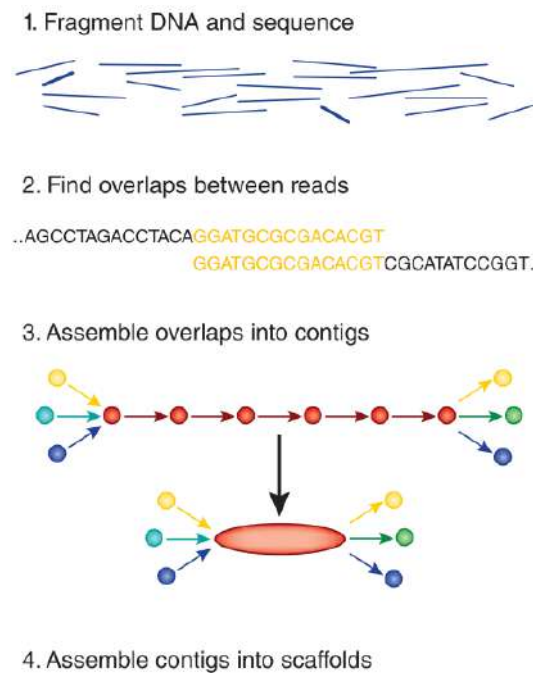


Fig. 2. El ensamblaje del genoma a partir de fragmentos cortos de ADN secuenciados (Baker, 2012)

Cuando se trata de especies modelo, como *Drosophila* o el ser humano, se utilizan los genomas de referencia, que facilitan el proceso de ensamblaje. A este proceso se le denomina mapeo o *mapping*. Los programas de *mapping* como BLAT o Tophat (éste último, diseñado específicamente para alinear secuencias

procedentes de RNA-Seq a un genoma de referencia para identificar uniones exón-exón; por lo general, usado en el pipeline: Bowtie > Tophat > Cufflinks y, opcionalmente con otras ampliaciones del paquete: Cuffcompare, Cuffmerge y Cuffdiff; Trapnell et al. 2009) requieren un genoma de referencia.

Cuando se carece del genoma de referencia, se recurre a ensambladores *de novo*. El estudio de Zhao et al. (2011) resulta especialmente interesante ya que compara diferentes ensambladores *de novo* y métodos de ensamblaje: (i) aquellos que utilizan un único k-mer (single k-mer, SK): SOAPdenovo, ABySS, Oases y Trinity, y (ii) aquellos que de estrategia múltiple k-mers (MK): SOAPdenovo-MK, trans-ABySS y Oases-MK (Fig 2). Las conclusiones de este estudio, junto a las otros adicionales, más el que se presenta en esta tesis, se discuten más adelante en la sección de Discusión.

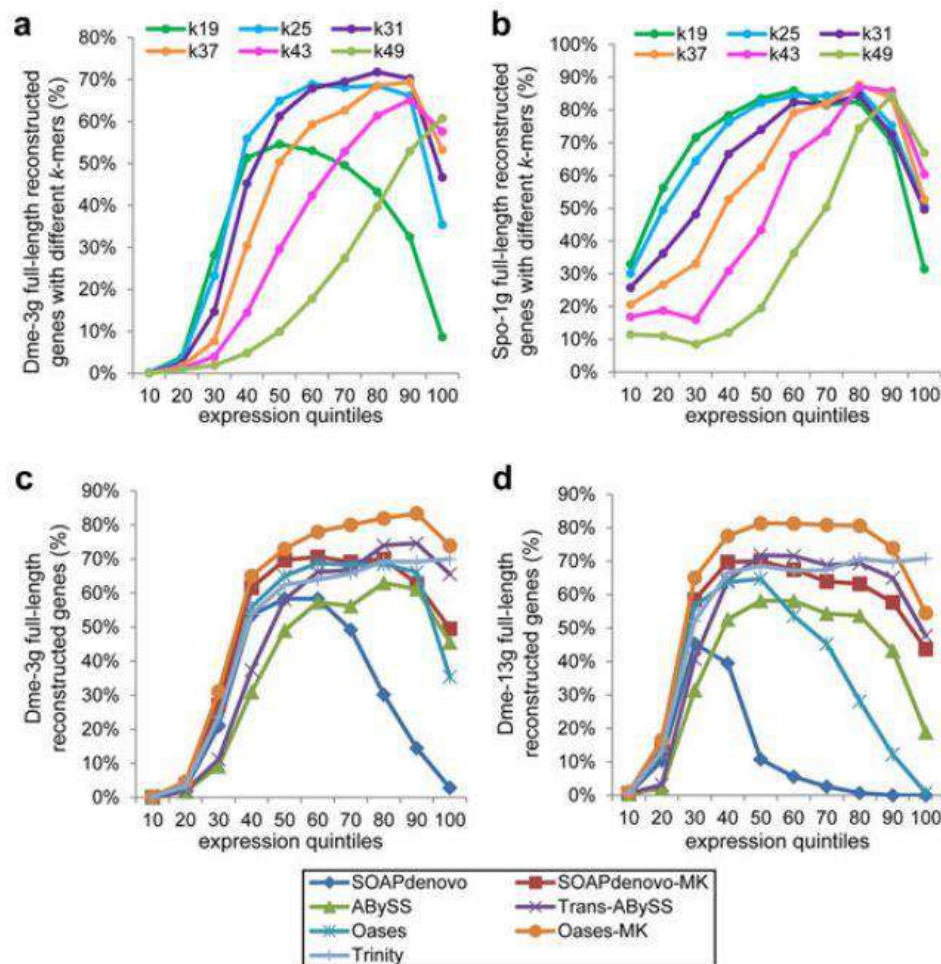


Fig. 3. Comparación de ensambladores *de novo* para el éxito en la reconstrucción de genes completos (Zhao et al. 2011)

En el panorama internacional, un esfuerzo considerable se ha llevado a cabo en el campo de la transcriptómica.

La mosca del vinagre, *Drosophila melanogaster*, es el insecto de estudio por excelencia y ha suministrado varios genomas de referencia y múltiples estudios transcriptómicos, pero no es así para el resto de especies de insectos.

En coleópteros, la aplicación de las tecnologías NGS mediante RNA-Seq de forma generalizada ya ha comenzado, ya sea con la secuenciación de tejidos particulares, como el tubo digestivo de *Chrysomela tremulae* (Pauchet et al. 2009) o la antena de *Anomala corpulenta* (Chen et al. 2014), entre otros, o de la secuenciación del transcriptoma de un espécimen completo, como en el caso del escarabajo del pino de montaña *Dendroctonus ponderosae* (Keeling et al. 2012) o el escarabajo de la patata *Leptinotarsa decemlineata* (Kumar et al. 2014).

Además, existen importantes trabajos de gran escala como i5k-Consortium (Hered, 2013), que pretende secuenciar 5.000 transcriptomas de artrópodos, o el proyecto 1KITE (<http://www.1kite.org/>, 2015), que incluye centros de investigación en EEUU, Alemania y China, y que aspira a secuenciar 1.000 transcriptomas de insectos.

Sin embargo y a pesar de lo anterior, el 90% de la diversidad de coleópteros está aún por categorizar (Benton et al. 2016) y huelga decir que más análisis son indudablemente necesarios.

#### 7.4. La expresión génica

La investigación transcriptómica consiste en caracterizar todos los transcritos presentes en una célula o tejido en un momento concreto y los mecanismos que dirigen su expresión (Ruan et al. 2004; Harbers & Carninci, 2005). Se encarga, pues, de secuenciar y analizar el conjunto de moléculas de mRNAs presentes en un momento de la vida celular.

El DNA que se transcribe (mRNA) es el único que interacciona fuera del núcleo celular con otras moléculas (receptores proteicos, enzimáticos, RNAs no mensajeros, etc.). Posee, por lo tanto, la capacidad de interferir en los procesos biológicos y puede, en consecuencia, determinar en mayor o menor medida la supervivencia del individuo.

Su condición fenotípica lo convierte, por fuerza, en un carácter supeditado a influir en el *fitness* (o éxito reproductivo) del individuo y, por lo tanto, a afectar a la selección natural a nivel poblacional.

Analizar la expresión génica es, por tanto, de ayuda a comprender cualquier proceso biológico y en cualquiera de sus niveles: molecular, fisiológico, individual, poblacional.

Comprendido el valor del mRNA frente al DNA en la expresión génica, no se ha dudado en estudiar su composición.

En la actualidad, contamos con excelentes plataformas para almacenar los datos de expresión génica.

La más antigua sea tal vez la actualmente conocida como The Uniprot Consortium (The Uniprot Consortium, 2017) que, aunque fundada en 2002 como la plataforma que conocemos hoy, procede de la unión de varios proyectos con una larga trayectoria en materia de análisis proteico. No es de extrañar que se nutra de proteínas, pues el primer elemento para medir la expresión génica fue precisamente la expresión proteica, debido al simple motivo de que la secuenciación de aminoácidos precedió a la secuenciación de nucleótidos.

No obstante, desde el momento en que se observó que ciertos RNAs no eran nunca traducidos a péptidos pero que, sin embargo, eran responsables de importantes funciones —algunas de ellas reguladoras—, el estudio de la expresión génica tuvo que poner su foco en el análisis del transcriptoma. Así, la

transcriptómica incluiría el análisis completo de todos los transcritos, algunos de los cuales son traducidos a polipéptidos y otros no, siendo todos interactores con moléculas de diversa índole. A continuación, véanse algunos ejemplos de RNAs: los fragmentos *non-codingRNAs* (ncRNAs) responsables de la traducción a proteínas (a modo de tRNA o rRNA), los microRNAs (miRNAs) involucrados en la regulación génica, los *long intergenic non-coding RNAs* (lincRNAs), los *small nuclear RNAs* (snRNAs), los *small nucleolar RNAs* (snoRNAs) y una lista creciente de tipos de ellos (Leung et al. 2013).

En cuanto a bases de datos de mRNA, contamos con FlyBase (Attril et al. 2016) que, entre otros, reúne los perfiles de expresión de genes individualizados de *D. melanogaster*, obtenidos mediante una variedad de procedimientos: (i) expresión transcriptómica deducida por *RT-PCR*, (ii) expresión polipeptídica ? obtenida mediante espectrometría de masas, (iii) expresión deducida de *reporters*, (iv) datos de expresión *high-throughput* mediante microarray y (v) datos de expresión *high-throughput* mediante *modEncode* RNA-Seq. Esta última reúne los datos de expresión génica que se deducen de los análisis de secuenciación con RNA-Seq.

Otra de las bases más conocidas en cuanto a expresión génica es FlyAtlas, la cual comprende 44 *arrays* de expresión de *Drosophila* (Affymetrix, Dros2), cada uno de ellos mapeando la expresión de 18.770 transcritos, lo que corresponde a la gran mayoría de genes de *Drosophila* conocidos ([http://flyatlas.org/about\\_atlas.html](http://flyatlas.org/about_atlas.html)). FlyAtlas está diseñado para encontrar rápidamente el mapa de expresión de un gen en cada tejido del adulto de *Drosophila*. Para ello, proporciona la información por gen y por tejido de: (i) la señal de mRNA (la abundancia de mRNA); (ii) el *enrichment* de mRNA (la cantidad de mRNA del tejido versus cuerpo entero); (iii) el Affymetrix *present call* (i.e., el número de veces detectado de cuatro *arrays* llevados a cabo).

Su contenido se ha ido actualizando y, a día de hoy, existe una nueva versión denominada FlyAtlas2.

Los datos recogidos en cuanto a la expresión génica deben ordenarse en un cuadro de función, que resuma y ayude a discernir la función/funciones de cada gen analizado. En este sentido, destaca el consorcio de ontología génica conocido como Gene Ontology Consortium (Gene Ontology Consortium, 2000), que nació en 1988 para la anotación funcional de los genes de los tres organismos modelo más utilizados: la mosca del vinagre *Drosophila melanogaster*, el ratón doméstico *Mus musculus* y la levadura *Saccharomyces cerevisiae*. Con posterioridad, dicho consorcio ha ido anexionando bases de datos adicionales de otras especies.



El Gene Ontology Consortium se divide en tres dominios fundamentales, que son la base de su clasificación interna: Biological Process (BP), Cellular Component (CC) y Molecular Function (MF), que hacen referencia a los tres grandes niveles de definición de funciones, a saber: funciones a nivel de los procesos biológicos del individuo, funciones a nivel de cada componente de la célula y funciones a nivel molecular de forma individualizada.

El dominio Biological Process (BP) comprende la colección de procesos que tienen lugar en los seres vivos, estratificándola en series jerarquizadas. De acuerdo con dicho consorcio, un proceso biológico se define como series reconocidas de eventos o funciones moleculares; además, para ser parte de la base de datos del BP, el término debe cumplir las condiciones de representar un proceso específico y un proceso entero (Gene Ontology Consortium, 2000).

## 7.5. Genes sex-biased en Insecta

La expresión diferencial de genes entre machos y hembras (a la cual los términos *sex-biased expression* y *sex-enriched expression* hacen alusión) ha sido analizada desde que se realizaron los primeros estudios de expresión.

Tradicionalmente, las diferencias entre ambos sexos se acusaban a la pertenencia o no de los cromosomas sexuales y a los genes contenidos en ellos (Singh & Jagadeeshan, 2012).

Efectivamente, algunos estudios indican que los genes sesgados de macho (*male-biased genes*) se encuentran infrarrepresentados en el cromosoma X, mientras que los genes sesgados de hembra (*female-biased genes*) se encuentran enriquecidos (Meisel et al. 2019; Meisel et al. 2012; Ellegren & Parsch, 2007; Parisi et al. 2003).

Sin embargo, hoy se conoce la existencia de una gran lista de genes que controlan una variedad de caracteres asociados a un sexo específico, ya sea en la categoría de *sex-biased* o *sex-enriched*, que serían los responsables del dimorfismo sexual (Singh & Jagadeeshan, 2012).

Hoy, gracias a las bases de datos mencionadas en el apartado anterior, que se nutren y ordenan la información procedente de los esfuerzos realizados en *Drosophila*, conocemos algunos datos de gran importancia para los científicos que analizan la expresión génica diferencial en insectos.

Además de las bases de datos, también existen estudios específicos.

Conocemos, por ejemplo, que aproximadamente el 30% del genoma de *Drosophila* muestra *sex-biased expression*, la mayoría en tejido reproductivo (Gravely et al. 2011; Ranz et al. 2003; Singh & Jagadeeshan, 2012).

Un estudio más reciente sobre *Drosophila* indica que la mayoría del genoma de ésta está sesgado sexualmente en algún momento de su ciclo vital, ya sea forma conservativa o mostrando un sesgo específico de alguna fase de su desarrollo (Perry et al. 2014).

Fuera del ámbito de estudio de *Drosophila*, Oppenheim et al. (2015) es consciente de dos trabajos solamente: (i) Baker et al. (2011), que encuentra que el 5-15% del genoma de *Anopheles gambiae* estaba sesgado sexualmente entre machos y hembras, y (ii) Wilkinson et al. (2013) que identifica ~900 *sex-biased genes* en moscas de ojos saltones (Diopsidae).

A expensas de Diptera, contamos, por ejemplo, con el análisis de Prince et al. (2010), el cual analiza el transcriptoma de *Tribolium castaneum* y encuentra que el ~20% de este está regulado diferencialmente entre machos y hembras. Además, identifica 416 ortólogos con expresión sexual sesgada en *T. castaneum*, *D.melanogaster*, y *A. gambiae*.

## 8. Objetivos

### *8.1. Objetivo general*

El objetivo general de la presente tesis es el análisis de la evolución molecular de manera comprensiva a través de la aplicación de las tecnologías más actuales —en especial la transcriptómica mediante RNA-Seq— de un grupo de elevada diversidad genética, el grupo taxonómico de los coleópteros, y sobre un tema de máximo interés como es la expresión sesgada de genes implicados en la reproducción.

## 8.2. Objetivos específicos

*Obtención de mRNA nuevo de muestras de coleópteros salvajes*

Secuenciar *de novo* cinco transcriptomas de cuatro especies no modelo, pertenecientes al género *Calligrapha* de crisomélidos (Coleoptera: Chrysomelidae), el cual posee diferentes estrategias reproductivas: *C. confluens*, *C. aff. floridana*, *C. multipunctata* y dos especímenes de *C. philadelphica*.

*Anotación funcional de novo de cinco transcriptomas de Calligrapha spp.*

Anotar funcionalmente *de novo* los recién secuenciados transcriptomas (de *C. confluens*, *C. aff. floridana*, *C. multipunctata* y *C. philadelphica*), utilizando los procedimientos más avanzados disponibles en la actualidad.

*Análisis descriptivo de los cinco transcriptomas por medio de comparaciones inter- e intraespecíficas*

Análisis de los contigs de cada transcriptoma específico mediante comparaciones intraespecíficas (entre *C. philadelphica* (PA) y *C. philadelphica* (QC)) e interespecíficas (entre el resto), de acuerdo a: la abundancia de contings, la similitud de las secuencias, la naturaleza de sus GO, la redundancia funcional de sus GO, etc.

Gene-finding

Examen de transcriptomas, mediante programación bioinformática, para el reclutamiento de genes de interés particular, en concreto para este estudio, de genes involucrados en funciones típicamente masculinas y, en especial, de genes responsables de la función de individualización de espermatozoides.

### *Análisis de evolución molecular*

Analizar la evolución molecular de cada uno de los 44 genes de individualización de espermatozoides (*sperm individualization*; GO:0007291) a lo largo de la clase Insecta: estimación de tasas evolutivas, identificación de ortólogos y de eventos de duplicación.

### *Análisis de la interacción génica*

Discernir la relación existente entre genes de individualización de espermatozoides mediante el análisis de posibles redes de interacción génica y la relación de éstas con la correspondiente tasa evolutiva.

## **9. Informe del factor de impacto de las publicaciones**



This PhD thesis include two research articles and both include statements about author contributions:

Article 1:

**Testis-specific RNA-Seq of *Calligrapha* (Chrysomelidae) as a transcriptomic resource for male-biased gene inquiry in Coleoptera**

**Vizán-Rico H.I., Gómez-Zurita J.**

*Mol. Ecol. Res.*, 17, 533-545 (2017)

doi:10.1111/1755-0998.12554

Journal Impact Factor: 7.059 (1st decile in "Biochemistry and Molecular Biology")

In this study, we functionally annotated five testis transcriptomes of four *Calligrapha* species and conducted similarity searches against male-biased genes in *Drosophila melanogaster* and *Tribolium castaneum* to shortlist the corresponding gene homologs in *Calligrapha*. Gene orthology and potential functional homology were confirmed for three of these genes based on phylogenetic analyses of gene trees. Helena's involvement in this work was profound, devoting most of her time to functionally annotate the testis transcriptomes, proposing informative ways to summarize the findings, and also performing dedicated similarity searches of sequence homologs in these transcriptomes against subsets of sex-biased genes in *Drosophila* and *Tribolium*.

Article 2:

**Patterns and constraints in the evolution of sperm individualization genes in insects, with an emphasis on beetles**

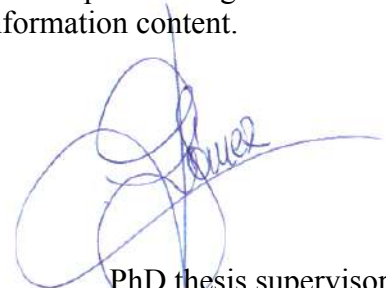
**Vizán-Rico H.I., Mayer C., Petersen M., McKenna D.D., Zhou X., Gómez-Zurita J.**

*Genes*, 10, 776 (2019)

doi:10.3390/genes10100776

Journal Impact Factor: 3.759 (2nd quartile in "Genetics and Heredity")

In this study, we selected a group of genes involved in reproductive processes in insects and analysed the patterns and processes affecting their evolution from a macroevolutionary perspective, with a particular interest in their duplication and evolutionary rates as well as potential correlates with known patterns of sex-biased expression. The analyses included the study of 44 genes involved in sperm individualization in insects and used homologous data from 119 insect model species and from 25 beetle species, 19 of them available from the 1KITE project. Twenty-one of these genes were identified as part of a single protein-protein and/or genetic/regulatory interaction network, and the evolutionary properties of these genes were also analysed in relation to the structure of this network. Helena contributed to this work by mining the ortholog sequence data for the beetle non-model species in the course of a short stay in the Alexander Koenig Museum in Bonn, where she learned how to use Orthograph for this task. She also performed preliminary multiple sequence alignments and phylogenetic analyses for all the genes in the study to help assess their information content.



PhD thesis supervisor:  
**Jesús Gómez-Zurita Frau**

## 10. Publicaciones

### 10.1. Publicación de primer autor nº1

Describimos la arquitectura genética de cinco transcriptomas de testículo de cuatro especies estrechamente relacionadas de escarabajos *Calligrapha* (Chrysomelidae), que divergieron durante los últimos 3 millones de años. Para ello, secuenciamos *de novo* cinco librerías de cDNA, utilizando Illumina HiSeq, obteniendo 102.884-176.514 contigs, y anotamos funcionalmente el ~33-45% de las que superaban 499 nt. Nuestras comparaciones interespecíficas de anotación y similitud de secuencia revelaron una alta homogeneidad en cuanto a composición génica, así como la presencia de varios candidatos funcionales involucrados en reproducción o procesos reproductivos (el 0,72-1,08% de las secuencias anotadas). Los estrictos análisis de similitud de secuencia entre los transcriptomas y un listado de genes de expresión sesgada masculina —demostrados empíricamente en *Drosophila melanogaster* y *Tribolium castaneum*— permitieron la identificación de 77 homólogos en las secuencias *Calligrapha*, posibles candidatos de expresión sesgada masculina. Para algunos de los cuales, incluyendo *CG9313*, *Tektin-A* y *tomboy40*, confirmamos la ortología con los genes de expresión sesgada masculina mediante inferencia filogenética y los datos de especies modelo de insecto disponibles, aumentando nuestra confianza en que también representan homólogos funcionales. Nuestros transcriptomas son un recurso transcriptómico valioso para el análisis de genes con sesgo de expresión masculino en *Calligrapha*, género que posee el interés adicional de incluir varias especies exclusivamente de hembras. Al mismo tiempo, representa un hito para estudios similares en Coleoptera, ampliando la diversidad taxonómica actualmente representada por la especie modelo *T. castaneum*, y datos genómicos incipientes en otros linajes herbívoros, incluidos gorgojos, cerambícidos y crisomélidos.

Vizán-Rico H.I., Gómez-Zurita J. (2017). Testis-specific RNA-Seq of *Calligrapha* (Chrysomelidae) as a transcriptomic resource for male-biased gene inquiry in Coleoptera. *Mol. Ecol. Res.*, 17, 533-545 (2017) doi:10.1111/1755-0998.12554

# Testis-specific RNA-Seq of *Calligrapha* (Chrysomelidae) as a transcriptomic resource for male-biased gene inquiry in Coleoptera

HELENA I. VIZÁN-RICO and JESÚS GÓMEZ-ZURITA

*Animal Biodiversity and Evolution, Institute of Evolutionary Biology (CSIC-Univ. Pompeu Fabra), Barcelona 08003, Spain*

## Abstract

We report the architecture of testis transcriptomes of four closely related species of *Calligrapha* (Chrysomelidae) beetles, which diverged during the last 3 million years. Five cDNA libraries were sequenced using Illumina HiSeq technology, retrieving 102 884–176 514 assembled contigs, of which ~33–45% of these longer than 499 nt were functionally annotated. Annotation and sequence similarity comparisons of these libraries revealed high homogeneity in gene composition and the presence of several functional candidates related to reproduction or reproductive processes (0.72–1.08% of annotated sequences). Stringent sequence similarity analyses of these transcriptomes against empirically demonstrated male-biased genes in *Drosophila melanogaster* and *Tribolium castaneum* allowed the identification of 77 homologues in *Calligrapha*, possible candidates of male-biased expression. Some of these genes – including CG9313, Tektin-A or tomboy40 – were confirmed as orthologs of these male-biased genes using phylogenetic inference and available model insect data, increasing our confidence that they represent functional homologues too. Our transcriptomes are a valuable transcriptomic resource for the analysis of male-biased genes in *Calligrapha*, which has the added interest of including several female-only species. But it simultaneously represents a landmark for similar studies in Coleoptera, broadening the taxonomic diversity currently represented by the model species *T. castaneum*, and incipient genomic data in other herbivorous lineages, including weevils, longhorn beetles and leaf beetles.

**Keywords:** candidate genes, functional annotation, gene orthology, nonmodel organisms, sex

Received 23 September 2015; revision received 23 May 2016; accepted 31 May 2016

## Introduction

The molecular control of phenotypic and specifically gender differences is a hot topic in biology. Despite males and females of any species share virtually all of their genes, phenotypic differences between both genders, from anatomical to physiological and behavioural traits, can be striking in some cases. Therefore, the explanation for sexual dimorphism must logically reside in the differences in gene expression profiles (Mank 2009; Assis *et al.* 2012; Parsch & Ellegren 2013). Genes that show these differential expression patterns are typically known as sex-biased genes, although there is no consensus on the thresholds that this differential expression should allude to (Assis *et al.* 2012). When this bias is extreme, up to a point when a gene is only expressed in gender-specific tissues or pathways, or with expression profile differences of several orders of magnitude, then it

is possible to talk about female- or male-specific genes (Parsch & Ellegren 2013). Finding these genes can be achieved with dedicated genetic and functional studies across tissues and developmental stages and there has been considerable progress in matching genes and functions in the case of model organisms (e.g. modENCODE database in the case of *Drosophila*; Graveley *et al.* 2011). However, results from one of these models cannot be easily extrapolated to other species, and this is particularly true for the expression of sex-biased genes, more in the case of male-biased genes, and genes involved in reproductive processes, which can be novel genes without easily recognizable homologues in other species, again particularly frequent in male-biased transcriptomes (Pröschel *et al.* 2006; Haerty *et al.* 2007; Assis *et al.* 2012; Ranz & Parsch 2012; Scolari *et al.* 2016). For non-model organisms, where the same level of experimentation is usually out of the question, a window of opportunity to gain insight on sex-biased genes and their expression is currently in place thanks to next-generation sequencing technologies. More specifically, when there is

Correspondence: Jesús Gómez-Zurita, Fax: +34 93 221 1011; E-mail: j.gomez-zurita@csic.es

no annotated genome available, which is the most common situation, *de novo* transcriptome assemblies from isolated mRNAs, the family of methods known as RNA-Seq (Martin & Wang 2011), provide valuable data to start digging into sex-biased genes and expression profiles. The most obvious approach that can be used to find sex-biased genes from RNA-Seq data is restricting the analysis to gender-specific tissues or organs, such as ovaries in the case of females or testicles and associated glands in the case of males, which enrich transcriptomes for genes known to be involved in sex determination and differentiation (e.g. Sun *et al.* 2013; Manousaki *et al.* 2014).

The affordability of these methods and their power have boosted an interest in the analysis of the molecular causes of phenotypic differences in nonmodel organisms based on RNA-Seq data, and sexual dimorphism has been one of the research targets, also in insects. Recently, Oppenheim *et al.* (2015) listed the handful of studies that have specifically examined the transcriptomes of male reproductive organs of nonmodel insects. These and other recent studies included the analysis of seminal proteins in *Heliconius* butterflies (Walters & Harrison 2010, 2011), mating-responsive genes in *Ceratitidis* Mediterranean fruit flies (Scolari *et al.* 2012), X-linked genes in stalk-eyed *Teleopsis* flies (Reinhardt *et al.* 2014), spermatogenesis genes (Wei *et al.* 2015) and testicular miRNAs (Tariq *et al.* 2016) in *Bactrocera* fruit flies, novel transcripts in the paternal sex ratio chromosome of *Nasonia* jewel wasps (Akbari *et al.* 2013), SSR development in *Periplaneta* American cockroach (Chen *et al.* 2015), or a broad view of transcriptome architecture in the phlebotomine sand fly *Lutzomyia* (Azevedo *et al.* 2012), the tsetse fly *Glossina* (Scolari *et al.* 2016), *Gryllus* and *Teleogryllus* field crickets (Andres *et al.* 2013; Bailey *et al.* 2013), or again in *Teleopsis* (Baker *et al.* 2012). To date, however, there are no similar studies in any representative of the huge Order Coleoptera.

We are interested in evolutionary aspects of the reproductive biology of the leaf beetle genus *Calligrapha*, whereby one of the most remarkable life history traits is the occurrence of several unisexual, female-only species (Robertson 1966; Gómez-Zurita *et al.* 2006; Montelongo & Gómez-Zurita 2015). In this case, it is intriguing to explore the evolutionary significance of males and male-specific functions and their eventual dispensability, at least in some species. This far-reaching question is nonetheless difficult to tackle experimentally, since *Calligrapha* leaf beetles, univoltine and strictly dependent on native North American trees (Brown 1945; Gómez-Zurita *et al.* 2006), are not laboratory-friendly organisms, and there are no genetic or genomic tools available to explore the molecular basis of maleness in this group. For these reasons, in order to start bridging the knowledge gap between phenotypic sex differences and sex exclusion and their molecular basis, we aim at providing here with

the first male tissue-specific transcriptomic resource for this group of beetles, which will in turn enrich available data for an insect group, Coleoptera, critically missing from similar approaches. Specifically, in this study, we have two distinctive goals: (i) characterizing the architecture of whole testis transcriptomes of several species of *Calligrapha* from a functional perspective and (ii) identifying potential male-biased genes in these transcriptomes by reference to functionally investigated transcriptomes of model insects.

## Materials and methods

### Source and preparation of samples

Male specimens of several species of *Calligrapha* were collected in their natural habitats in a single entomological campaign in northeastern North America during the late spring of 2012. Specimens were killed upon capture, dissected to remove abdominal tergites for better exposition to fixatives and preserved in RNAlater tissue storage reagent (Sigma-Aldrich Química SL, Madrid) before storage in the laboratory at  $-80^{\circ}\text{C}$ . Five adult males were selected for study, including two of *Calligrapha philadelphia* (L.) (USA, PA, Gifford Pinchot Pk.; and Canada, PQ, Cascapédia-St. Jules) and one each of *C. confluens* Schaeffer (Canada, NB, Miramichi), *C. aff. floridana* Schaeffer (USA, MD, Patuxent River Pk.) and *C. multipunctata* (Say) (USA, PA, Allegheny Natl. Ft.). Testes were dissected from the specimens in RNAlater, individually kept in vials with fresh RNAlater, stored with dry ice and submitted for RNA extraction.

### cDNA library preparation and sequencing

Library preparation was outsourced to the NGS laboratory of Eurofins MWG Operon (Ebersberg, Germany). Total RNA was extracted from each individual pair of testicles with yields averaging  $6.1 \pm 2.12 \mu\text{g}$  (ranging from  $2.4 \mu\text{g}$  in *C. aff. floridana* to  $7.5 \mu\text{g}$  in *C. multipunctata*). Total RNA quantity and quality as assessed using MultiNA microchip electrophoresis (Shimadzu Corp., Kyoto, Japan) were suitable for cDNA synthesis. All the available RNA from each individual sample was used to create short insert (150–400 bp) cDNA libraries customized for high-throughput Illumina HiSeq 2000 (Illumina Inc., San Diego CA) sequencing. Each library was obtained by polyA purification and fractionation of RNA, double-strand cDNA synthesis using random priming and end repair, ligation of Illumina sequencing and 6-mer indexing adaptors, purification of fragments and PCR amplification to generate the sequencing library, with necessary quality controls. Pools of individually indexed libraries were sequenced in three channels of

Illumina HiSeq 2000 using a shotgun protocol with chemistry v3.0 and 2 × 100 bp paired-end read modules.

### Quality check and de novo assembly

Sequences were assigned to individual libraries based on their index code and forbidding any mismatch, adaptor sequences were trimmed and resulting raw Illumina reads were filtered for quality. Quality clipping removed low-quality ends (Phred score < 20) and also clipped reads below the same quality threshold assessed using a sliding window approach (window size = 4). Short reads, below 70 bp, were removed too. Raw data cleaning and trimming were performed with *trimmomatic* 0.22 (Lohse *et al.* 2012). *De novo* assembly of quality-filtered data was achieved by means of a bioinformatic pipeline based on contig assembly using a multi-k-mer approach by *VELVET* 1.2.08 (Zerbino & Birney 2008) assisted by *GraphConstructor* (Convey Computer Corp., Richardson TX), followed by contig clustering of splice variants (isotigs) using *OASES* 0.2.08 (Schulz *et al.* 2012) and of isogroups, ideally genes, using *NEWBLER* 2.6 (Roche Diagnostics Corp., Indianapolis IN). Default parameters were used in every case. Assembled contigs obtained from each library were deposited in the European Nucleotide Archive database (EBI-EMBL, Hinxton, UK) under the study Accession no. PRJEB13133 and sequence Accession nos. HADL01000001–HADL01028570 (*C. confluens*), HADL01028571–HADL01060532 (*C. floridana*), HADL01060533–HADL01089880 (*C. multipunctata*), HADL01089881–HADL01115995 (*C. philadelphica* from PA) and HADL01115996–HADL01144017 (*C. philadelphica* from PQ).

### Functional annotation

*De novo* functional annotation of assembled contigs in *Calligrapha* was performed using the standard approach implemented in *BLAST2GO* (Conesa *et al.* 2005). *BLAST2GO* submits query sequences to similarity searches using *BLAST* (Altschul *et al.* 1990), extracting the Gene Ontology (GO; Gene Ontology Consortium, 2000) or other functional annotation terms of the corresponding hits when these have been characterized functionally, and attaching these functional labels to the query sequences. Functional annotation was restricted to contigs longer than 499 nt. Contigs were analysed against NCBI's nonredundant protein database using the translated *BLAST*, *blastx*, algorithm and a relatively low similarity cut-off threshold (E-value = 10E-3) and the obtained results were submitted to the mapping step using default options. Annotation was improved in every case using refinements based on several InterProScan 5 tools and their default parameters for identification of protein domains (Jones *et al.* 2014) and the ENZYME (Bairoch 2000) database. GO terms derived

from *Calligrapha* testes were organized according to the three general hierarchical functional categories, namely biological processes (BP), molecular functions (MF) and cellular components (CC), and up to the second ontogeny level of the multilevel combined graph generated by *BLAST2GO* for each general category. Taxonomic information attached to mapping results was analysed for the presence of bacteria and viruses with a customized script counting matches in mapping annotation files with 3639 genera (6 April 2016) obtained from relevant taxID in GenBank (taxalD: Bacteria = 2, Archaea = 2157, Viruses = 10239).

The output of *BLAST2GO* is more easily analysed unlinked from the sequence data that generated it, because the number of retrieved GO terms is much higher than the number of available contigs (each contig is typically associated with several such terms) and GO terms do not conform to a strict nested hierarchy (the same GO terms can be in several categories at different levels). The classification of these terms is thus preferably done using semantic criteria. We used *GOOSE*, the SQL Environment of the Gene Ontology database of the *AMIGO* 2 (Carbon *et al.* 2009) online search tool, to extract the list of all child terms from a Gene Ontology node of interest. The functionally annotated transcriptomes of *Calligrapha* were analysed for matches with these specific GOs using iterative string recognition commands in Perl.

### Redundancy analysis of *Calligrapha* transcriptomes

GO term redundancy from each library and among libraries was compared also using a custom Perl script for term recognition. Moreover, as an exploratory alternative to detect compositional discrepancies between libraries, we evaluated the overlap between annotated genes of library pairs (one acting as experimental set, and the other as source of GO annotated genes) using the two-tailed Fisher's exact test (FET) with a relaxed false discovery rate (FDR) < 0.05 in *BLAST2GO*. The underlying idea of this exploration of global overlap is that libraries fundamentally different in architecture (and associated annotations) will render a high number of significant FET results. Finally, similarity or redundancy between libraries was also assessed in terms of nucleotide sequence similarity by estimating the proportion of contigs larger than 499 nt in one library producing at least one significant hit against another library using local *BLAST* searches with the *blastn* algorithm and stringent similarity criteria (E-value = 10E-30).

### Classification based on similarity searches (of functional studies)

The collection of *Calligrapha* testis isotigs, reassembled contigs recognized as splicing gene variants from the



more stringent NEWBLER assembly procedure and longer than 499 nt was investigated to find their homologues among collections of genes from model insects which have been related to male function. The first source of male-biased insect genes for comparison was the wealth of functional and gene expression studies carried out in *Drosophila* and curated in FlyBase (release February 24th, 2015; Dos Santos *et al.* 2015). In particular, we filtered the modENCODE expression by stage data (Graveley *et al.* 2011), selecting exclusively those genes not expressed or with extremely low expression in adult females (0 RPKM as defined by FlyBase) and overexpressed (from high to extremely high expression; 51 to >1000 RPKM) in adult males. The second source of genes was the report of Prince *et al.* (2010) on male-biased genes in the flour beetle *Tribolium castaneum* (Herbst) linked to the X chromosome. Even if this study was not extensive to the whole genome, it is an extremely valuable resource for the purpose of our research since it is one of the few, if not the only study of this nature conducted on another beetle, and sexual chromosomes tend to accumulate nonetheless genes of sex-related functions (Vicoso & Charlesworth 2006; Bellott *et al.* 2010; Brelsford *et al.* 2013; but see Meisel *et al.* 2012). The genes with the expression profile of interest reported in this study were extracted from the supplementary material of the article and downloaded from the latest release of BeetleBase (Kim *et al.* 2010). The translated CDSs of the genes selected from *Drosophila* and *Tribolium* were used to find potential homologues in each one of the testis transcriptomes of *Calligrapha*. Homology searches used in every case the *blastx* algorithm applying a stringent E-value = 10E-30.

#### *Structural and phylogenetic homology and hypotheses of functionality*

*Calligrapha* genes with structural homologues of male-biased genes in *Tribolium* were subject to a new round of stringent (E-value = 10E-30) *blastx* searches against the collection of male-biased polypeptides in *Drosophila* with structural homologues in *Calligrapha*. The obtained hits represented *Calligrapha* genes with male-biased expression profiles in two other holometabolous insects, thus potential candidates to be overexpressed in *Calligrapha* males as well. These sequences were subject to functional annotation with BLAST2GO as above, which supported the selection of a reduced number of candidate genes for further analyses. These sequences with their BLAST2GO annotations were deposited in the European Nucleotide Archive database (EBI-EMBL, Hinxton, UK) under Accession nos. HADM01000001–HADM01002329. Structural and phylogenetic homology of these genes was assessed estimating their phylogenetic position among insect orthologs. Prior to phylogenetic analyses,

*Calligrapha* transcript sequences were translated to amino acids using the standard genetic code in GENEIOUS 8.0.3 (Biomatters Ltd.) and the ORF matching the known functional protein was extracted for subsequent analyses. Ortholog searches used text-based queries based on the identifier of the gene of interest in *Drosophila* using the online resource OrthoDB v8 (Kriventseva *et al.* 2015) and applying a taxonomic filter to retrieve only orthologs among the 80 insect species with complete genomes and covered by this database. Besides *Tribolium*, the OrthoDB database includes three other beetles increasingly closer to *Calligrapha*, namely *Dendroctonus ponderosae* Hopkins (Curculionidae), *Anoplophora glabripennis* (Motschulsky) (Cerambycidae) and *Leptinotarsa decemlineata* (Say) (Chrysomelidae). Multiple sequence alignment of insect sequences with *Calligrapha* orthologs was done using default parameters of the G-INS-i algorithm and the BLOSUM62 scoring matrix in MAFFT 7 (Katoh & Standley 2013). Alignments were inspected for obvious problematic sequences, which were excluded before inferring maximum-likelihood trees and bootstrap support based on 100 pseudoreplicates, using default parameters and the LG amino acid substitution model (Le & Gascuel 2008) in PHYML (Guindon & Gascuel 2003). Trees were rooted secondarily on Palaeopteran branches, represented by *Ephemera danica* Müller (Ephemeroptera), *Ladona fulva* (Müller) (Odonata) or both.

## Results

### *cDNA libraries of Calligrapha testes*

The first challenge for our experimental work was obtaining RNA from specimens collected in the wild, far away from the laboratory, and with sufficient quantity and quality for cDNA library preparation. Particularly, worrisome was the need to work with testicles from a single male specimen, representing a tiny amount of tissue. The results from the RNA isolation and cDNA library synthesis showed that our collection methods and the amount of tissue available were clearly sufficient for highly satisfactory results. Sequence yield and quality of the 2 × 100 bp paired-end Illumina sequencing were comparable for all libraries, slightly improved for *C. aff. floridana*, which had lower amount of starting RNA for library synthesis (Table 1). Illumina runs produced between 13 and over 16 000 Mbp of sequence data of which an average of 98.13% for all libraries survived quality control checks. These data resulted in high number of contigs in every case, ranging between 102 882 in *C. philadelphica* from Quebec and 176 514 in *C. aff. floridana*, the latter producing the longest assembled contig (23 942 nt). The amount of contigs longer than 0.5 kb was well balanced among libraries, ranging from 18.11%

**Table 1** Comparative analysis of the characteristics of five testis-specific cDNA libraries obtained from four species of *Calligrapha*

Testis library	Yield (Mbp)	Surviving reads*	Total contigs	Total length	Max. length	Contigs >499 nt	Isotigs >499 nt	Max. length	%GC	%N
<i>C. confluens</i>	12 984	127 201 453 (97.97%)	120 369	54 631 334	15 499	28 570	41 568	27 374	39.9	0.0025
<i>C. aff. floridana</i>	15 668	153 817 718 (98.17%)	176 514	65 986 846	23 942	31 962	48 295	60 815	39.9	0.0020
<i>C. multipunctata</i>	13 025	127 745 967 (98.08%)	141 513	55 069 928	14 963	29 348	44 571	21 406	39.4	0.0025
<i>C. philadelphica</i> PA	16 202	159 095 526 (98.20%)	117 755	49 178 211	14 761	26 115	39 420	24 666	40.0	0.0022
<i>C. philadelphica</i> QC	14 026	137 758 393 (98.21%)	102 882	51 083 549	12 761	28 022	42 595	21 827	40.0	0.0018

\*Number of reads retained after quality check (Phred score < 20) with trimmomatic 0.22 (Lohse *et al.* 2012).

in *C. aff. floridana* to 27.24% in *C. philadelphica* from Quebec. Finally, libraries yielded a high number of assembled splice variants of genes larger than 0.5 kb, from 39 420 isotigs in *C. philadelphica* from Pennsylvania to 48 295 in *C. aff. floridana*, all with negligible ambiguity in base calls and representing an average GC content of 39.84% (Table 1).

#### Functional annotation of cDNA libraries

Functional *de novo* annotation of *Calligrapha* testis libraries against the nr database expectedly retrieved most (81.34–83.51%) of top hits among the two species of beetles with most of annotated Coleoptera sequence data in public sequence databases. Approximately 2/3 of these hits were with the flour beetle *Tribolium castaneum*, and the remainder with a bark beetle, the mountain pine beetle *Dendroctonus ponderosae*. With a dramatic drop in frequency, some top hits were still attributed to three other insect model species in three orders including, in decreasing order, the pea aphid *Acyrtosiphon pisum* Harris (Hemiptera), the jewel wasp *Nasonia vitripennis* (Walker) (Hymenoptera) and the silk moth *Bombyx mori* (Linnaeus) (Lepidoptera). Occasional top hits were found for a large number of animals, typically those for which some genomic work exists. The content of the libraries was assessed for the presence of prokaryote and viral sequences to inform on the quality of source eukaryotic RNA. On average, each library contained 1.13% of prokaryote or viral matches (from 0.49% in *C. philadelphica* from Pennsylvania to 1.72% in *C. multipunctata*). These matches were mostly attributed to Eubacteria, and the vast majority (55.4% on average for both *C. philadelphica* samples, and 70.9% for the others) specifically on genes of *Wolbachia*, known endosymbiont of all North American species of *Calligrapha* analysed so far (J. Gómez-Zurita, unpublished).

Annotation success was relatively high for contigs longer than 0.5 kb in all *Calligrapha* libraries. The proportion of annotated sequences ranged between 32.8% in *C. aff. floridana* and 44.6% in *C. philadelphica* from Pennsylvania (Table 2). Annotations represented an ensemble

of 37 377–48 881 extracted GO terms depending on the library and identified 4670–5596 unique functions. The largest proportion of annotated contigs (86.1–88.6%) were assigned to the molecular function domain either on its own or combined with terms spanning additional domains, while other combinations accounted for 66.0–71.7% of annotated contigs in the case of biological process domain and for 43.1–50.0% in the cellular component domain. Most mapping matches (89.95–92.03%) were contributed by the UniProt Knowledgebase (UniProt EMBL-EBI/SIB/PIR Consortium) and their annotations mainly (69.95–74.67% of sequences) inferred from evidence provided by automated methods without curatorial judgement (IEA category, with default Evidence Code = 0.7; Gene Ontology Consortium).

Table S1 (Supporting information) shows the selection of GO terms included in the second level of the directed acyclic graph for the three GO domains in each of the *Calligrapha* testis libraries. The functional composition of the five libraries was very similar, with enrichment analyses of pairwise comparisons of libraries producing non-significant results for Fisher's exact tests in all cases except in comparisons with the library of *C. floridana*. In this case, the test produced significant results with over-representation in the library of *C. floridana* for relatively few GO terms, ranging between four terms when *C. confluens* was the test library and up to 38 terms when it was *C. philadelphica* from Quebec (relative to *C. philadelphica* from Pennsylvania, there were 36 significant results, including the only three terms under-represented in the library *C. floridana*). In all libraries, there was a clear dominance (77.7–81.3%) of genes related to cellular, metabolic, single-organism and biological regulation processes, to cellular components, including membranes, organelles and macromolecular complexes, and to binding and catalytic activities. Functional representation in the libraries was rather homogeneous too if considering the full Gene Ontology hierarchy, whereby up to 67.6% of all GO terms of any ontology level were shared between at least two libraries (or 52.7% if redundancy was removed). We additionally examined the homogeneity of libraries based on sequence similarity criteria,



**Table 2** Functional annotation of assembled contigs >499 nt in five testis-specific cDNA libraries from *Calligrapha*, and distribution of annotated contigs in the highest hierarchical levels

Testis library	Annotated contigs	Assigned GOs	Unique GOs	BP-CC-MF	BP-MF	MF	CC-MF	BP-CC	CC	BP
<i>C. confluens</i>	11 409	40 187	4670	3350	3319	2816	533	509	527	355
<i>C. aff. floridana</i>	10 490	43 618	5037	3748	2949	1939	495	469	530	360
<i>C. multipunctata</i>	10 423	37 377	4696	3122	2897	2488	467	499	592	358
<i>C. philadelphica</i> PA	11 656	48 881	5596	4062	2990	2440	544	542	572	367
<i>C. philadelphica</i> QC	11 038	39 141	4723	3249	3223	2730	578	548	549	336

BP, biological process; MF, molecular function; and CC, cellular component of Gene Ontology (GO) mapping.

**Table 3** Number of isotigs longer than 499 nt in testis-specific *Calligrapha* cDNA libraries producing at least one significant *blastn* hit (E-value 10E-30) in the pairwise comparison of libraries

Libraries	Query Lib.	FLO	MUL	PHI <sub>QC</sub>	PHI <sub>PA</sub>
Search Lib.	No. contigs >499 nt	48 295	44 571	42 595	39 420
CON	41 568	38 210	36 705	38 800	36 685
FLO	48 295	—	41 526	39 107	41 229
MUL	44 571	—	—	37 741	36 641
PHI <sub>QC</sub>	42 595	—	—	—	38 605

CON, *C. confluens*; FLO, *C. aff. floridana*; MUL, *C. multipunctata*; PHI, *C. philadelphica* from Pennsylvania and Quebec.

and interlibrary comparisons suggested high sequence redundancy, even though these results are inflated to some extent, because of genes represented by several contigs in one library having homologues in another library also represented by several contigs. Significant hits between libraries ranged from 85.04% between the libraries of *C. confluens* and *C. floridana* to 94.14% between the two libraries of *C. philadelphica*, representing on average 89.05% of shared isotigs longer than 499 nt among libraries (Table 3).

#### Retrieval of functional data related to reproduction

The most obvious GO terms to look for male functions potentially represented in testis cDNA libraries were two descendant terms from the biological process category, namely 'reproductive process' (GO:0022414) and 'reproduction' (GO:0000003). *Calligrapha* testis libraries contained 183–318 and 180–329 functional annotations falling within these two categories, respectively (Table S1, Supporting information). These terms subtend 824 and 934 child terms, respectively, or 1234 unique terms when both are combined, of which 184 were represented in the testis libraries of *Calligrapha* (Table S2, Supporting information). Most (44.0%) of these GO terms were present in all or four of the libraries, but a high

proportion (15.8%) was privative of one library. Focusing on the analysis of splicing variants of genes longer than 499 bp, most of the annotated functions were associated with female reproductive physiology or female germ line function (35.3%) as well as with meiotic division (18.5%), but a number of annotations (20.7% of 902 annotated contigs mapping on descendants of the focal GO nodes) were annotated with male-specific functions, either in male meiosis, spermatogenesis, development or behaviour (Table 4). These male-function-related GO terms were found in all or four of the libraries in 40.5% of the cases, and 37.8% were exclusive of one library. Functional annotations with more contigs mapping onto them were also found in all libraries without exception, and they included functions such as male meiosis cytokinesis (GO:0007112), imaginal disc-derived male genitalia morphogenesis (GO:0048803), male courtship behaviour (GO:0008049), sperm individualization (GO:0007291), spermatid development (GO:0007286), spermatocyte division (GO:0048137) and spermatogenesis (GO:0007283).

#### Similarity searches against *Drosophila* and *Tribolium* male-biased genes

A reference database of male-biased genes from *Drosophila* included 538 genes, represented by 746 different polypeptides, which were queried using *blastx* with splice variant gene sequences longer than 499 nt from each of the testis libraries of *Calligrapha*. Table 5 shows the proportion of *Calligrapha* genes finding a match with male-biased *Drosophila* genes using a high-stringency similarity search criterion. Stringent *BLAST* searches resulted in 1.78–2.32% of *Calligrapha* testis library genes finding homologous sequences among 207–211 male-biased genes of *Drosophila* (representing 213 different loci). Out of 213 male-biased *Drosophila* polypeptides with at least one homologue in *Calligrapha*, a core of 200 was shared among all *Calligrapha* libraries (Table S3, Supporting information). In turn, we isolated 1310 different polypeptides from the X chromosome in *Tribolium* identified by Prince *et al.* (2010) as overexpressed in males.

**Table 4** Distribution of functional annotation terms with roles suggestive of male specificity and based on the Gene Ontology hierarchy for assembled splicing variants of genes >499 nt in five cDNA libraries obtained from testis of four *Calligrapha* species

Process GO-id	GO term	CON	FLO	MUL	PHI <sub>PA</sub>	PHI <sub>QC</sub>
Isotigs mapping to reproduction or reproductive process		141	210	137	140	274
Unique GO-id		110	129	92	144	120
Male meiosis						
GO:0007053	Spindle assembly involved in male meiosis	1	1	—	2	1
GO:0007054	Spindle assembly involved in male meiosis I	1	1	1	1	1
GO:0007060	Male meiosis chromosome segregation	—	1	—	1	1
GO:0007112	Male meiosis cytokinesis	1	10	4	5	3
GO:0007140	Male meiosis	1	4	—	3	1
GO:0007141	Male meiosis I	2	1	1	1	1
Gametogenesis						
GO:0007283	Spermatogenesis	13	10	11	18	13
GO:0007285	Primary spermatocyte growth	1	—	1	1	—
GO:0007286	Spermatid development	6	4	4	14	7
GO:0007287	Nebenkern assembly	—	—	—	1	1
GO:0007289	Spermatid nucleus differentiation	—	—	1	—	1
GO:0007290	Spermatid nucleus elongation	—	1	1	1	1
GO:0007291	Sperm individualization	6	10	8	16	8
GO:0048137	Spermatocyte division	2	1	5	5	3
Male development						
GO:0007485	Imaginal disc-derived male genitalia development	3	1	—	2	2
GO:0030238	Male sex determination	—	1	—	—	—
GO:0030539	Male genitalia development	1	—	1	1	—
GO:0030724	Testicular fusome organization	—	—	—	2	—
GO:0030850	Prostate gland development	—	—	—	1	—
GO:0048092	Negative regulation of male pigmentation	1	3	—	—	—
GO:0048803	Imaginal disc-derived male genitalia morphogenesis	4	6	3	6	5
GO:0048808	Male genitalia morphogenesis	—	—	—	—	1*
GO:0060442	Branching involved in prostate gland morphogenesis	—	—	1	—	—
GO:0060516	Primary prostatic bud elongation	—	—	—	—	1*
GO:0060523	Prostate epithelial cord elongation	—	—	—	—	1*
GO:0060685	Regulation of prostatic bud formation	—	—	—	—	1*
GO:0060740	Prostate gland epithelium morphogenesis	—	1	—	—	—
GO:0060769	Positive regulation of epithelial cell proliferation involved in prostate gland development	—	—	—	—	1*
GO:0060770	Negative regulation of epithelial cell proliferation involved in prostate gland development	—	—	—	1	1
GO:0060782	Regulation of mesenchymal cell proliferation involved in prostate gland development	—	—	—	—	1*
GO:0060783	Mesenchymal smoothed signalling pathway involved in prostate gland development	—	—	—	—	1*
Male behaviour						
GO:0008049	Male courtship behaviour	10	4	2	8	5
GO:0042628	Mating plug formation	1	1	—	—	—
GO:0042713	Sperm ejaculation	—	—	1	—	—
GO:0045433	Male courtship behaviour, veined wing generated song production	1	3	1	1	1
GO:0060179	Male mating behaviour	—	1	1	2	2
GO:0060406	Positive regulation of penile erection	—	—	—	—	1*

CON, *C. confluens*; FLO, *C. aff. floridana*; MUL, *C. multipunctata*; PHI, *C. philadelphia* from Pennsylvania and Quebec.

\*These annotations related to 'prostate' development and regulation are alternative annotations for the same contig.

The stringent similarity searches against these flour beetle genes found 915 candidate genes on average in the *Tribolium* genome and for a relatively high proportion of tested *Calligrapha* sequences (12.36–16.38%). The average

length of these hits was above 400 nt and with very high similarity (Table 5).

A crossed similarity analysis between positive hits simultaneously for *Drosophila* and *Tribolium* retrieved

**Table 5** Results of *blastx* similarity searches of assembled genes >499 nt in five cDNA libraries obtained from testis of four *Calligrapha* species against the databases of male-biased genes in *Drosophila melanogaster* and X-linked male-biased genes in *Tribolium castaneum*. Searches used a threshold of similarity based on a stringent E-value (E-30)

	Isotigs	Hits	Isotigs with match	Matching loci	Alignment length (SD)	Average E-value	Bit score (SD)
<i>Drosophila</i> male-biased							
<i>C. confluens</i>	41 568	1907	965	207	309.50 ± 117.99	6.60E-33	223.83 ± 143.48
<i>C. floridana</i>	48 295	1863	1019	208	325.01 ± 126.31	9.54E-33	247.80 ± 169.28
<i>C. multipunctata</i>	44 571	1406	789	208	317.25 ± 125.32	1.64E-32	242.59 ± 150.25
<i>C. philadelphia</i> PA	39 420	1692	900	211	315.57 ± 130.63	1.32E-32	232.95 ± 160.57
<i>C. philadelphia</i> QC	42 595	1899	956	209	316.91 ± 121.30	9.49E-33	236.95 ± 146.78
<i>Tribolium</i> male-biased							
<i>C. confluens</i>	41 568	14 978	6208	916	455.58 ± 460.57	1.56E-32	311.07 ± 369.66
<i>C. floridana</i>	48 295	14 841	6599	917	404.49 ± 393.52	1.78E-32	299.09 ± 333.93
<i>C. multipunctata</i>	44 571	12 543	5511	911	416.52 ± 464.51	1.26E-32	312.14 ± 408.12
<i>C. philadelphia</i> PA	39 420	14 476	6138	917	450.24 ± 499.31	1.37E-32	331.29 ± 430.90
<i>C. philadelphia</i> QC	42 595	16 108	6978	915	438.50 ± 509.56	1.55E-32	325.76 ± 415.43

SD, standard deviation.

3533 isotigs longer than 499 nt among the five libraries (from 606 in *C. multipunctata* to 761 in *C. confluens*) that matched up to 124 male-biased polypeptides (genes and alternative inferred splicing) in *Drosophila* and *Tribolium*. These sequences represented in some cases different translation patterns for a total of 92 genes, of which 77 had been successfully characterized in the four species of *Calligrapha* (Table S3, Supporting information). Functional annotation of these genes, the candidates for male-biased expression also in *Calligrapha*, is shown in Table S4 (Supporting information).

## Discussion

### *Suitability of experimental procedure*

RNA-Seq based on the Illumina platform has become the method of choice for two large classes of biological enquiry (Van Verk *et al.* 2013): analysis of gene expression and gene finding, which is our current area of interest. Gene finding benefits from experimental designs aiming at an enrichment of these genes supposedly associated with the relevant process, which may have a very low representation in the transcriptome under scrutiny (Robles *et al.* 2012). Specifically, it may be advantageous to apply sample preparation techniques either to remove unwanted functions or to target specific ones. Among the first, hybridization subtraction libraries remove unnecessary functions by combining cDNA from the tissue of interest with cDNA from another tissue or physiological condition chemically tagged for subsequent removal upon hybridization with the former (e.g. Olsvik *et al.* 2013). The study of male-specific testis function

could subtract female germ line transcripts, or even whole female transcriptomes, which would likely reduce the number of the gender-unspecific and housekeeping functions reported in Table S1 (Supporting information). Other existing methodologies are better at retrieving specific genes with low or differential expression, such as the targeted RNA sequencing method known as capture sequencing (CaptureSeq; Mercer *et al.* 2014), although the results produced would not be amenable to differential expression analyses.

Nonetheless, dealing with nonmodel organisms, for which no genomic, genetic or expression data are available, a shotgun Illumina-based approach is still highly efficient and provides with additional advantages. In our approach, we intentionally aimed at characterizing a snapshot of the whole testis transcriptome, including all functions gender unspecific. Robust *de novo* transcriptome assembly thanks to reliable bioinformatic tools, even if affected by some biases, still yields a huge number of putative transcripts. This yield of data is big enough to include mRNAs in low representation together with abundant data on other genes, which, even if temporarily uninteresting for the purpose of a particular study, represent a cost-effective strategy allowing for other uses of data. Admittedly, our shotgun approach generated a volume of information far greater than needed or useful for the original purpose of the study, so that only 0.38–0.75% of isotigs longer than 499 nt seemingly received functional annotations suggestive of the targeted process, male specificity. Nonetheless, this small percentage still represented up to 37 recognized male functions (Table 4), and by virtue of the uniformity across library composition (suggesting that tissue

processing and library preparation and analysis issues were not problematic in this case), these male functions and associated genes were available in most cases for all *Calligrapha* species investigated here.

Tissue-specific shotgun approaches may raise other concerns in RNA-Seq experiments such as the accuracy in extraction and processing of tissue (Johnson *et al.* 2013), sequencing-related issues such as depth of coverage or transcript length biases or bioinformatic limitations for the correct detection of isoforms. However, particularly the first three concerns are more problematic in quantitative applications, not so much for gene discovery (Oppenheim *et al.* 2015). In part contributing to these biases, our libraries were not normalized, which may be considered a drawback reducing the efficiency for gene finding, particularly in the case of these with lower levels of expression. However, again invoking the power of big numbers allowed by Illumina technology, this strategy proves advantageous providing with a quantitative flavour to data (Moghadam *et al.* 2013), tentatively exploited here to assess the uniformity of the five *Calligrapha* libraries, but eventually useful to compare expression levels.

#### RNA-Seq and male-biased gene annotation

Most functional annotation studies using RNA-Seq data report a large proportion, above 50% and up to 80% or more, of assembled transcripts lacking annotation (e.g. Wang *et al.* 2010; Riesgo *et al.* 2012). These same proportions are typical of transcriptome analyses of gonadal tissues, with annotation success ranging 22.5–38.5% (Sun *et al.* 2013; Chen *et al.* 2015; Meng *et al.* 2015). In our case, annotation success reached 32.8–44.6%, but these figures may be comparatively inflated, since we restricted annotation to contigs longer than 499 nt, with higher annotation rates compared to shorter fragments (Riesgo *et al.* 2012). Moreover, annotation success depends on the efficiency of the initial BLAST step, so that functional annotation based on results of the *blastx* algorithm, as used here, consistently outperforms the same based on the *blastn* algorithm (e.g. Vidotto *et al.* 2013; Lamanna *et al.* 2014). Despite the high annotation success for *Calligrapha* testis libraries, there are still a large proportion of putative transcripts left out based on functional annotation analyses. Male-biased genes are reported in many cases as fast-evolving, neo-functionalized genes that are often short (Ellegren & Parsch 2007; Haerty *et al.* 2007; Magnusson *et al.* 2011). These characteristics may compromise the ability of BLAST-based strategies to find them, and we acknowledge that a large proportion of effectively male-biased genes may remain in the nonannotated fraction of the *Calligrapha* libraries, because of lack of overall similarity and also size culling in our case. In our libraries, whereas the fractions of contigs longer than 1000 nt and

contigs of 500–999 nt approximately contained the same number of sequences (ratio of longer vs. shorter fraction: 0.90–1.02), the latter contributed a notably lower proportion of successfully annotated contigs (26.5% vs. 51.8%, on average). Indeed, dedicated studies targeting precisely these novel genes in testis transcriptomes have shown that they may actually represent a significant fraction of deep sequencing data (e.g. Akbari *et al.* 2013; Scolari *et al.* 2016). In any case, in these initial stages of genomic enquiry for a nonmodel system, reference to known genes is already a huge step forward in building knowledge for further research and our approach still represents extremely useful for gene finding (Oppenheim *et al.* 2015).

There are several studies that investigate the testis transcriptome in a variety of organisms focusing on specific genes or processes (e.g. Akbari *et al.* 2013; Chen *et al.* 2015; Meng *et al.* 2015). Yet, the global comparison of testis transcriptomes, unless targeting close relatives or physiological alternatives in the same species, may be pointless, given that each transcriptome reflects not only taxonomic, but also individual physiological idiosyncrasy (Scolari *et al.* 2016). An interesting comparison, however, affects the proportion of annotations which are specifically related to reproduction and reproductive processes, which in the case of *Calligrapha* reached an ensemble 0.72–1.08%, similar, but slightly lower to values obtained in other organisms, typically around 1.5% and reaching up to 2% of annotations (e.g. He *et al.* 2012; Gao *et al.* 2014; Chen *et al.* 2015).

#### Candidate male-biased genes in *Calligrapha*

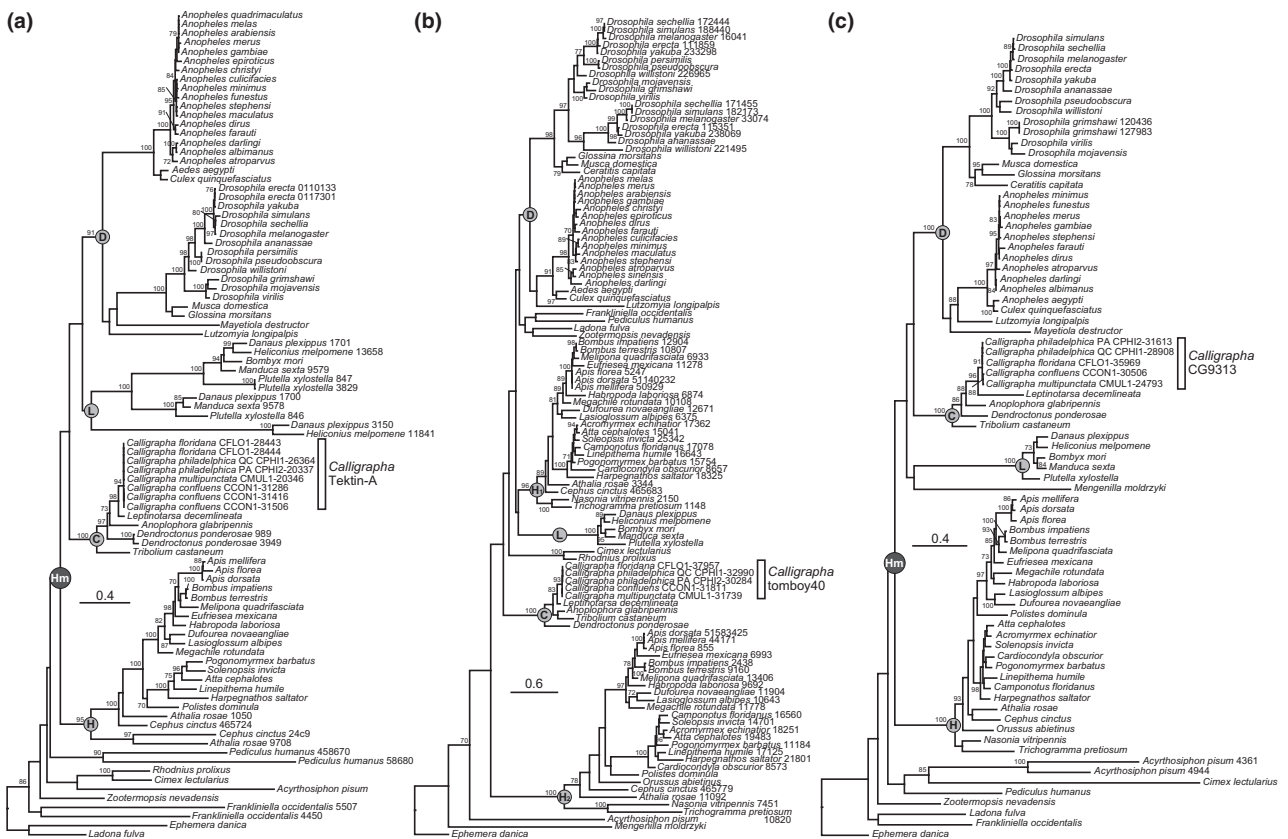
*Calligrapha* testis libraries contributed between 606 and 761 transcripts homologous to 124 male-biased genes both in *Drosophila* and *Tribolium*. With every precaution, we hypothesize that these genes may be equally male-biased in *Calligrapha*. These precautions are needed because it has been shown that a high proportion of male-biased genes in one insect species may fail to find orthologs in relatives of the same Order (Scolari *et al.* 2016). Thus, our cross-taxa sequence similarity search strategy may incur biases due to exclusion of male-biased novel *Calligrapha* genes and inclusion of male-biased genes in *Drosophila* and *Tribolium* which, nonetheless, are not differentially expressed in *Calligrapha*.

Out of these 124 genes, a total of 77 have been isolated in the four species of *Calligrapha* investigated here and this shortlist constitutes the core of candidate genes for future functional and evolutionary studies of male-specific functions in these and related organisms (Table S3, Supporting information). Among these, we highlight these homologous to *Drosophila* genes CG9313, CG10859, Tektin-A, kinesins Klp59C and Klp59D, all involved in flagellar motility (Rogers *et al.* 2004; Dorus *et al.* 2006;



Graveley *et al.* 2011; Robinson *et al.* 2013); the glycogen synthase kinase-3 known as *gskt*, involved in male gamete generation and male gonad development (Kalamegham *et al.* 2007; Dos Santos *et al.* 2015); *tomboy40*, a transmembrane transporter specific of insect male germ lines (Hwa *et al.* 2004); or *TTL3B*, which has been related to sperm individualization (Rogowski *et al.* 2009). The orthology of *Calligrapha* genes was assessed through phylogenetic analyses including several insects (e.g. Table S5, Supporting information). The selected genes and obtained trees showed some properties of interest making them strong suitable candidates for the study of the evolution of male function and specificity in insects, but most clearly in the case of Coleoptera (see examples in Fig. 1). In general, the polypeptides inferred from these genes show relatively high conservation at the amino acid level even for the whole Class and more clearly within holometabolans. At the same time, nucleotide sequence disparity among closely related species is not negligible (e.g. *p*-distances of 0–0.011 [*tomboy40*],

0.005–0.013 [CG9313] or 0.009–0.029 [Tektin-A] among *Calligrapha* species separated in the past 2–3.5 Ma; Montelongo & Gómez-Zurita 2014). Moreover, these genes exhibit discrete paralogy, sometimes affecting entire Orders or lineages within Orders (e.g. two copies of *tomboy40* in both Hymenoptera and *Drosophila* among Diptera, or three copies of Tektin-like genes in Lepidoptera; Fig. 1). However, at least for the genes analysed in detail, there is no evidence of paralogy affecting Coleoptera. Some of these trees, even if poorly supported at basal branching events, retain phylogenetic structure reflecting the current systematics of insects. At lower divergence levels, the expected phylogenetic relationships are obtained with high support in some cases, most remarkably in the Coleoptera clade, but also among dipterans and hymenopterans, relatively well sampled. These properties combined can facilitate, for instance, primer design targeting different taxonomic ranks in evolutionary studies of insects (e.g. Wild & Maddison 2008). Moreover, even though the specific function that



**Fig. 1** Maximum-likelihood phylogenetic trees and bootstrap support based on inferred amino acid sequences of three male-biased genes in insects, including Tektin-A (a), *tomboy40* (b) and CG9313 (c). The position of the orthologs newly characterized in this work from testis-specific cDNA libraries from several species of *Calligrapha* is indicated with a white box, and monophyletic taxonomic groups consistent with insect systematics are highlighted too (C: Coleoptera; D: Diptera; H: Hymenoptera; Hm: Holometabola (=Endopterygota); and L: Lepidoptera). Trees rooted secondarily on the branch leading to *Ephemera danica*, as representative of the infra-class Paleoptera.

these genes undertake in *Calligrapha* is completely unknown, having characterized them from structural and phylogenetic perspectives makes them amenable for experimentation in our and other related nonmodel species.

## Acknowledgements

Dr. A. Santure (University of Auckland, New Zealand) and two anonymous reviewers contributed many constructive suggestions that helped us to improve several weaknesses in our manuscript. Tinguaro Montelongo (IBE, Barcelona) helped us in the field to obtain fresh samples of *Calligrapha* for transcriptomic analyses. This research was funded by project CGL2011-23820 of the Spanish Ministry of Economy and Competitiveness to JGZ, and an associated predoctoral studentship of the FPI programme to HIVR.

## References

- Akbari OS, Antoshechkin I, Hay BA, Ferree PM (2013) Transcriptome profiling of *Nasonia vitripennis* testis reveals novel transcripts expressed from the selfish B chromosome, paternal sex ratio. *Genes, Genomes, Genetics*, **3**, 1597–1605.
- Altschul SF, Gish W, Miller W, Myers EW, Lipman DJ (1990) Basic local alignment search tool. *Journal of Molecular Biology*, **215**, 403–410.
- Andres JA, Larson EL, Bogdanowicz SM, Harrison RG (2013) Patterns of transcriptome divergence in the male accessory gland of two closely related species of field crickets. *Genetics*, **193**, 501–513.
- Assis R, Zhou Q, Bachtrog D (2012) Sex-biased transcriptome evolution in *Drosophila*. *Genome Biology and Evolution*, **4**, 1189–1200.
- Azevedo RV, Dias DB, Bretas JA *et al.* (2012) The transcriptome of *Lutzomyia longipalpis* (Diptera: Psychodidae) male reproductive organs. *PLoS ONE*, **7**, e34495.
- Bailey NW, Veltsos P, Tan YF, Millar AH, Ritchie MG, Simmons LW (2013) Tissue-specific transcriptomics in the field cricket *Teleogryllus oceanicus*. *Genes, Genomes, Genetics*, **3**, 225–230.
- Bairoch A (2000) The ENZYME database in 2000. *Nucleic Acids Research*, **28**, 304–305.
- Baker RH, Narechania A, Johns PM, Wilkinson GS (2012) Gene duplication, tissue-specific gene expression and sexual conflict in stalk-eyed flies (Diopsidae). *Philosophical Transactions of the Royal Society B*, **367**, 2357–2375.
- Bellott DW, Skaletsky H, Pyntikova T *et al.* (2010) Convergent evolution of chicken Z and human X chromosomes by expansion and gene acquisition. *Nature*, **466**, 612–616.
- Brelsford A, Stöck M, Betto-Colliard C *et al.* (2013) Homologous sex chromosomes in three deeply divergent anuran species. *Evolution*, **67**, 2434–2440.
- Brown WJ (1945) Food-plants and distribution of the species of *Calligrapha* in Canada, with description of new species (Coleoptera, Chrysomelidae). *Canadian Entomologist*, **77**, 117–133.
- Carbon S, Ireland A, Mungall CJ, Shu S, Marshall B, Lewis S (2009) AmiGO: online access to ontology and annotation data. *Bioinformatics*, **25**, 288–289.
- Chen W, Liu Y-X, Jiang G-F (2015) *De novo* assembly and characterization of the testis transcriptome and development of EST-SSR markers in the cockroach *Periplaneta americana*. *Scientific Reports*, **5**, 11144.
- Conesa A, Götz S, García-Gómez JM, Terol J, Talón M, Robles M (2005) Blast2GO: a universal tool for annotation, visualization and analysis in functional genomics research. *Bioinformatics*, **21**, 3674–3676.
- Dorus S, Busby SA, Gerike U, Shabanowitz J, Hunt DF, Karr TL (2006) Genomic and functional evolution of the *Drosophila melanogaster* sperm proteome. *Nature Genetics*, **38**, 1440–1445.
- Dos Santos G, Schroeder AJ, Goodman JL *et al.* (2015) FlyBase: introduction of the *Drosophila melanogaster* Release 6 reference genome assembly and large-scale migration of genome annotations. *Nucleic Acids Research*, **43**, D690–D697.
- Ellegren H, Parsch J (2007) The evolution of sex-biased genes and sex-biased gene expression. *Nature Reviews Genetics*, **8**, 689–698.
- Gao J, Wang Xw, Zou Zh, Jia Xw, Wang Yl, Zhang Zp (2014) Transcriptome analysis of the differences in gene expression between testis and ovary in green mud crab. *BMC Genomics*, **15**, 585.
- Gene Ontology Consortium (2000) Gene ontology: tool for the unification of biology. *Nature Genetics*, **25**, 25–29.
- Gómez-Zurita J, Funk DJ, Vogler AP (2006) The evolution of unisexuality in *Calligrapha* leaf beetles: molecular and ecological insights on multiple origins via interspecific hybridization. *Evolution*, **60**, 328–347.
- Graveley BR, Brooks AN, Carlson JW *et al.* (2011) The developmental transcriptome of *Drosophila melanogaster*. *Nature*, **471**, 473–479.
- Guindon S, Gascuel O (2003) A simple, fast, and accurate algorithm to estimate large phylogenies by maximum likelihood. *Systematic Biology*, **52**, 696–704.
- Haerty W, Jagadeeshan S, Kulathinal RJ *et al.* (2007) Evolution in the fast lane: rapidly evolving sex-related genes in *Drosophila*. *Genetics*, **177**, 1321–1335.
- He L, Wang Q, Xk Jin *et al.* (2012) Transcriptome profiling of testis during sexual maturation stages in *Eriocheir sinensis* using Illumina sequencing. *PLoS ONE*, **7**, e33735.
- Hwa JJ, Zhu AJ, Hiller MA, Kon CY, Fuller MT, Santel A (2004) Germline specific variants of components of the mitochondrial outer membrane import machinery in *Drosophila*. *FEBS Letters*, **572**, 141–146.
- Johnson BR, Atallah J, Plachetzki DC (2013) The importance of tissue specificity for RNA-seq: highlighting the errors of composite structure extractions. *BMC Genomics*, **14**, 586.
- Jones P, Binns D, Chang H-Y *et al.* (2014) InterProScan 5: genome-scale protein function classification. *Bioinformatics*, **30**, 1236–1240.
- Kalamegham R, Sturgill D, Siegfried E, Oliver B (2007) *Drosophila* *mojiless*, a retroposed GSK-3, has functionally diverged to acquire an essential role in male fertility. *Molecular Biology and Evolution*, **24**, 732–742.
- Katoh K, Standley DM (2013) MAFFT multiple sequence alignment software version 7: improvements in performance and usability. *Molecular Biology and Evolution*, **30**, 772–780.
- Kim HS, Murphy T, Xia J *et al.* (2010) BeetleBase in 2010: revisions to provide comprehensive genomic information for *Tribolium castaneum*. *Nucleic Acids Research*, **38**, D437–D442.
- Kriventseva EV, Tegenfeldt F, Petty TJ *et al.* (2015) OrthoDB v8: update of the hierarchical catalog of orthologs and the underlying free software. *Nucleic Acids Research*, **43**, D250–D256.
- Lamanna F, Kirschbaum F, Tiedemann R (2014) *De novo* assembly and characterization of the skeletal muscle and electric organ transcriptomes of the African weakly electric fish *Campylomormyrus compressirostris* (Mormyridae, Teleostei). *Molecular Ecology Resources*, **14**, 1222–1230.
- Le SQ, Gascuel O (2008) An improved general amino-acid replacement matrix. *Molecular Biology and Evolution*, **25**, 1307–1320.
- Lohse M, Bolger AM, Nagel A *et al.* (2012) RobiNA: a user-friendly, integrated software solution for RNA-Seq-based transcriptomics. *Nucleic Acids Research*, **40**, W622–W627.
- Magnusson K, Mendes AM, Windbichler N *et al.* (2011) Transcription regulation of sex-biased genes during ontogeny in the malaria vector *Anopheles gambiae*. *PLoS ONE*, **6**, e21572.
- Mank JE (2009) Sex chromosomes and the evolution of sexual dimorphism: lessons from the genome. *American Naturalist*, **173**, 141–150.
- Manousaki T, Tsakogiannis A, Lagnel J *et al.* (2014) The sex-specific transcriptome of the hermaphrodite sparid sharpnose seabream (*Diplodus puntazzo*). *BMC Genomics*, **15**, 655.
- Martin JA, Wang Z (2011) Next-generation transcriptome assembly. *Nature Reviews Genetics*, **12**, 671–682.
- Meisel RP, Malone JH, Clark AG (2012) Disentangling the relationship between sex-biased gene expression and X-linkage. *Genome Research*, **22**, 1255–1265.

- Meng X-I, Liu P, Jia F-I, Li J, Gao B-Q (2015) *De novo* transcriptome analysis of *Portunus trituberculatus* ovary and testis by RNA-Seq: identification of genes involved in gonadal development. *PLoS ONE*, **10**, e0128659.
- Mercer TR, Clark MB, Crawford J *et al.* (2014) Targeted sequencing for gene discovery and quantification using RNA CaptureSeq. *Nature Protocols*, **9**, 989–1009.
- Moghadam HK, Harrison PW, Zachar G, Székely T, Mank JE (2013) The plover neurotranscriptome assembly: transcriptomic analysis in an ecological model species without reference genome. *Molecular Ecology Resources*, **13**, 696–705.
- Montelongo T, Gómez-Zurita J (2014) Multilocus molecular systematics and evolution in time and space of *Calligrapha* (Coleoptera: Chrysomelidae, Chrysomelinae). *Zoologica Scripta*, **43**, 605–628.
- Montelongo T, Gómez-Zurita J (2015) Non-random patterns of genetic admixture expose the complex historical hybrid origin of unisexual leaf beetle species in the genus *Calligrapha*. *American Naturalist*, **185**, 113–134.
- Olsvik PA, Berg V, Lyche JL (2013) Transcriptional profiling in burbot (*Lota lota*) from Lake Mjøsa – a Norwegian lake contaminated by several organic pollutants. *Ecotoxicology and Environmental Safety*, **92**, 94–103.
- Oppenheim SJ, Baker RH, Simon S, DeSalle R (2015) We can't all be supermodels: the value of comparative transcriptomics to the study of non-model insects. *Insect Molecular Biology*, **24**, 139–154.
- Parsch J, Ellegren H (2013) The evolutionary causes and consequences of sex-biased gene expression. *Nature Reviews Genetics*, **14**, 83–87.
- Prince EG, Kirkland D, Demuth JP (2010) Hyperexpression of the X chromosome in both sexes results in extensive female bias of X-linked genes in the flour beetle. *Genome Biology and Evolution*, **2**, 336–346.
- Pröschel M, Zhang Z, Parsch J (2006) Widespread adaptive evolution of *Drosophila* genes with sex-biased expression. *Genetics*, **174**, 893–900.
- Ranz JM, Parsch J (2012) Newly evolved genes: moving from comparative genomics to functional studies in model systems. How important is genetic novelty for species adaptation and diversification?. *BioEssays*, **34**, 477–483.
- Reinhardt JA, Brand CL, Paczolt KA, Johns PM, Baker RH, Wilkinson GS (2014) Meiotic drive impacts expression and evolution of x-linked genes in stalk-eyed flies. *PLoS Genetics*, **10**, e1004362.
- Riesgo A, Andrade SCS, Sharma PP *et al.* (2012) Comparative description of ten transcriptomes of newly sequenced invertebrates and efficiency estimation of genomic sampling in non-model taxa. *Frontiers in Zoology*, **9**, 33.
- Robertson JG (1966) The chromosomes of bisexual and parthenogenetic species of *Calligrapha* (Coleoptera: Chrysomelidae) with notes on sex ratio, abundance and egg number. *Canadian Journal of Genetics and Cytology*, **8**, 695–732.
- Robinson SW, Herzyk P, Dow JA, Leader DP (2013) FlyAtlas: database of gene expression in the tissues of *Drosophila melanogaster*. *Nucleic Acids Research*, **41**, D744–D750.
- Robles JA, Qureshi SE, Stephen SJ, Wilson SR, Burden CJ, Taylor JM (2012) Efficient experimental design and analysis strategies for the detection of differential expression using RNA-Sequencing. *BMC Genomics*, **13**, 484.
- Rogers GC, Rogers SL, Schwimmer TA *et al.* (2004) Two mitotic kinesins cooperate to drive sister chromatid separation during anaphase. *Nature*, **427**, 364–370.
- Rogowski K, Juge F, van Dijk J *et al.* (2009) Evolutionary divergence of enzymatic mechanisms for posttranslational polyglycylation. *Cell*, **137**, 1076–1087.
- Schulz MH, Zerbino DR, Vingron M, Birney E (2012) Oases: robust *de novo* RNA-seq assembly across the dynamic range of expression levels. *Bioinformatics*, **28**, 1086–1092.
- Scolari F, Gomulski LM, Ribeiro JM *et al.* (2012) Transcriptional profiles of mating responsive genes from testes and male accessory glands of the Mediterranean fruit fly, *Ceratitis capitata*. *PLoS ONE*, **7**, e46812.
- Scolari F, Benoit JB, Michalkova V *et al.* (2016) The spermatophore in *Glossina morsitans morsitans*: insights into male contributions to reproduction. *Scientific Reports*, **6**, 20334.
- Sun Fy, Liu Sk, Gao Xy *et al.* (2013) Male-biased genes in catfish as revealed by RNA-Seq analysis of the testis transcriptome. *PLoS ONE*, **8**, e68452.
- Tariq K, Peng W, Saccone G, Zhang H (2016) Identification, characterization and target gene analysis of testicular microRNAs in the oriental fruit fly *Bactrocera dorsalis*. *Insect Molecular Biology*, **25**, 32–43.
- Van Verk MC, Hickman R, Pieterse CM, Van Wees SC (2013) RNA-Seq: revelation of the messengers. *Trends in Plant Science*, **18**, 175–179.
- Vicoso B, Charlesworth B (2006) Evolution on the X chromosome: unusual patterns and processes. *Nature Reviews Genetics*, **7**, 645–653.
- Vidotto M, Grapputo A, Boscarì E *et al.* (2013) Transcriptome sequencing and *de novo* annotation of the critically endangered Adriatic sturgeon. *BMC Genomics*, **14**, 407.
- Walters JR, Harrison RG (2010) Combined EST and proteomic analysis identifies rapidly evolving seminal fluid proteins in *Heliconius* butterflies. *Molecular Biology and Evolution*, **27**, 2000–2013.
- Walters JR, Harrison RG (2011) Decoupling of rapid and adaptive evolution among seminal fluid proteins in *Heliconius* butterflies with divergent mating systems. *Evolution*, **65**, 2855–2871.
- Wang X-W, Luan J-B, Li J-M, Bao Y-Y, Zhang Ch-X, Liu S-S (2010) *De novo* characterization of a whitefly transcriptome and analysis of its gene expression during development. *BMC Genomics*, **11**, 400.
- Wei D, Li H-M, Yang W-J *et al.* (2015) Transcriptome profiling of the testis reveals genes involved in spermatogenesis and marker discovery in the oriental fruit fly, *Bactrocera dorsalis*. *Insect Molecular Biology*, **24**, 41–57.
- Wild AL, Maddison DR (2008) Evaluating nuclear protein-coding genes for phylogenetic utility in beetles. *Molecular Phylogenetics and Evolution*, **48**, 877–891.
- Zerbino DR, Birney E (2008) Velvet: algorithms for *de novo* short read assembly using de Bruijn graphs. *Genome Research*, **18**, 821–829.

---

J.G.-Z. designed the study. H.I.V.R. and J.G.-Z. collected the specimens for the study, and J.G.-Z. identified, prepared and dissected the samples. Generation of transcriptomic data was outsourced. H.I.V.R. and J.G.-Z. analysed and interpreted the data. J.G.-Z. wrote the manuscript and both authors approved its final version.

---

### Data accessibility

Raw data and assembled contigs obtained from each library were deposited in the European Nucleotide Archive database (EBI-EMBL, Hinxton, UK) under the primary Accession no. PRJEB13133. Contigs obtained via VELVET assembly were given Accession nos. HADL01000001–HADL01028570 (*C. confluens*), HADL01028571–HADL01060532 (*C. floridana*), HADL01060533–HADL01089880 (*C. multipunctata*), HADL01089881–HADL01115995 (*C. philadelphica* from PA) and HADL01115996–HADL01144017 (*C. philadelphica* from PQ). Sequences corresponding to the set of candidate genes with putative male-biased expression have been deposited with their corresponding annotation under Accession nos. HADM01000001–HADM01002329.

## Supporting Information

Additional Supporting Information may be found in the online version of this article:

**Table S1** Distribution of functional annotation terms in the second level of the Gene Ontology hierarchy for assembled contigs >499 nt in five cDNA libraries obtained from testis of four *Calligrapha* species (CON: *C. confluens*; FLO: *C. aff. floridana*; MUL: *C. multipunctata*; PHI: *C. philadelphica* from Pennsylvania and Quebec)

**Table S2** Gene Ontology (GO) functional annotation terms children to Reproduction (GO:0000003) and Reproductive Process (GO:0022414) found in the testis-specific cDNA libraries of five species of *Calligrapha* (CON: *C. confluens*; FLO: *C. aff. floridana*; MUL: *C. multipunctata*; PHI: *C. philadelphica* from Pennsylvania and Quebec)

**Table S3** Collection of male-biased *Drosophila melanogaster* polypeptides extracted from modENCODE expression-by-stage data and filtered to retrieve genes not expressed or with extremely low expression in adult females, over-expressed (high or extremely high expression) in adult males, and retrieved in at least one of the testis-specific cDNA libraries of five species of *Calligrapha* (CON: *C. confluens*; FLO: *C. aff. floridana*; MUL: *C. multipunctata*; PHI: *C. philadelphica* from Pennsylvania and Quebec)

**Table S4** Distribution of functional annotation terms in the second level of the Gene Ontology hierarchy for assembled isotigs >499 nt in five cDNA libraries obtained from testis of four *Calligrapha* species (CON: *C. confluens*; FLO: *C. aff. floridana*; MUL: *C. multipunctata*; PHI: *C. philadelphica* from Pennsylvania and Quebec)

**Table S5** Insect taxa and gene IDs used to infer the insect phylogenies of genes CG9313, Tektin-A and tomboy40 and their orthologs in *Calligrapha* as shown in Fig. 1 of the main article



## 10.2. Publicación de primer autor nº2

Los perfiles de expresión génica pueden cambiar radicalmente entre sexos, y el sesgo sexual puede contribuir a una dinámica macroevolutiva específica para los genes de expresión sexual sesgada. Sin embargo, estas dinámicas no se comprenden bien a gran escala evolutiva, debido a la escasez de estudios que han evaluado la ortología y la homología funcional para genes de expresión sexual sesgada, así como a los efectos pleiotrópicos que posiblemente están limitando su potencial evolutivo. En este trabajo, exploramos la correlación de la expresión sexualmente sesgada con los procesos macroevolutivos asociados a los genes sexualmente sesgados, incluidas las duplicaciones y las tasas evolutivas aceleradas. En concreto, examinamos dichos rasgos en un grupo de 44 genes que orquestan la individualización de los espermatozoides durante el proceso biológico de la espermatogénesis, con una expresión tan imparcial como sexualmente sesgada. El estudio se realiza en el amplio marco evolutivo de la clase Insecta, con especial atención en el orden Coleoptera. Utilizamos 119 genomas de insectos (incluyendo 6 especies modelo de Coleoptera) y 19 transcriptomas adicionales de especies de colópteros. Para el subconjunto de proteínas que interactúan física y/o genéticamente, también analizamos cómo su estructura de red puede condicionar el modo de evolución genética. El conjunto de genes fue muy heterogéneo en cuanto a duplicaciones, tasas de evolución y la estabilidad de la tasa, pero hubo evidencia estadística de correlación entre el sesgo sexual y tasas de evolución más rápidas, en consonancia con las predicciones teóricas. Las tasas más rápidas también mostraron correlación con patrones de sustitución ajustados a un reloj molecular (en secuencias aminoacídicas de insectos) y no ajustadas (en nucleótidos de coleópteros) en dichos genes. Las asociaciones estadísticas (tasas más altas para los nodos centrales) o la falta de las mismas (centralidad de genes duplicados) contrastaban con algunas hipótesis evolutivas actuales, lo que destaca la necesidad de más investigación sobre el presente tema.

Vizán-Rico H.I., Mayer C., Petersen M., McKenna D.D., Zhou X., Gómez-Zurita J. (2017). Patterns and constraints in the evolution of sperm individualization genes in insects, with an emphasis on beetles. *Genes*, 10, 776 (2019) doi:10.3390/genes10100776

Article

# Patterns and Constraints in the Evolution of Sperm Individualization Genes in Insects, with an Emphasis on Beetles

Helena I. Vizán-Rico <sup>1</sup>, Christoph Mayer <sup>2</sup> , Malte Petersen <sup>2</sup> , Duane D. McKenna <sup>3</sup> ,  
Xin Zhou <sup>4</sup> and Jesús Gómez-Zurita <sup>1,\*</sup> 

<sup>1</sup> Animal Biodiversity and Evolution, Institute of Evolutionary Biology (CSIC-Universitat Pompeu Fabra), 08003 Barcelona, Spain; helena.vizan@ibe.upf-csic.es

<sup>2</sup> Center for Molecular Biodiversity Research, Zoological Research Museum Alexander Koenig, 53113 Bonn, Germany; c.mayer.zfmk@uni-bonn.de (C.M.); malte.petersen@senckenberg.de (M.P.)

<sup>3</sup> Center for Biodiversity Research, Department of Biological Sciences, University of Memphis, Memphis, TN 38152, USA; dmckenna@memphis.edu

<sup>4</sup> Department of Entomology, College of Plant Protection, China Agricultural University, Beijing 100193, China; xinzhoucaddis@icloud.com

\* Correspondence: j.gomez-zurita@csic.es; Tel.: +34-93-2309643

Received: 24 August 2019; Accepted: 1 October 2019; Published: 4 October 2019



**Abstract:** Gene expression profiles can change dramatically between sexes and sex bias may contribute specific macroevolutionary dynamics for sex-biased genes. However, these dynamics are poorly understood at large evolutionary scales due to the paucity of studies that have assessed orthology and functional homology for sex-biased genes and the pleiotropic effects possibly constraining their evolutionary potential. Here, we explore the correlation of sex-biased expression with macroevolutionary processes that are associated with sex-biased genes, including duplications and accelerated evolutionary rates. Specifically, we examined these traits in a group of 44 genes that orchestrate sperm individualization during spermatogenesis, with both unbiased and sex-biased expression. We studied these genes in the broad evolutionary framework of the Insecta, with a particular focus on beetles (order Coleoptera). We studied data mined from 119 insect genomes, including 6 beetle models, and from 19 additional beetle transcriptomes. For the subset of physically and/or genetically interacting proteins, we also analyzed how their network structure may condition the mode of gene evolution. The collection of genes was highly heterogeneous in duplication status, evolutionary rates, and rate stability, but there was statistical evidence for sex bias correlated with faster evolutionary rates, consistent with theoretical predictions. Faster rates were also correlated with clocklike (insect amino acids) and non-clocklike (beetle nucleotides) substitution patterns in these genes. Statistical associations (higher rates for central nodes) or lack thereof (centrality of duplicated genes) were in contrast to some current evolutionary hypotheses, highlighting the need for more research on these topics.

**Keywords:** Coleoptera; evolutionary rates; gene network; Insecta; phylogenetic inference; sex-biased genes

## 1. Introduction

Phenotypic and physiological differences among closely related species with highly similar genomes are expected to be the result of differences in the expression profiles of key genes (e.g., [1]). In this regard, understanding the mechanisms underlying differences between males and females of the same species becomes of particular interest. Conspecific individuals of different sexes share most, if

not all, of their genome and genetics but sometimes display striking anatomical and physiological differences. Studies using model organisms have demonstrated the existence of significant differences in gene expression profiles between sexes. For example, approximately 30% of genes in the vinegar fly (order Diptera), *Drosophila melanogaster*, show sex-biased expression, and most of these genes are specific to reproductive tissues [2–4]. In fact, it has been proposed that most gene expression in *Drosophila* is sex biased at some point, exhibiting this bias either throughout the life cycle or in specific developmental stages [5]. Similarly, 5–15% of the genes in the mosquito (order Diptera) *Anopheles gambiae* genome show differential expression between males and females [6], and approximately 20% of the X-chromosome genes of *Tribolium castaneum* (order Coleoptera) are regulated differently in each sex [7].

The existence and need for biases in gene expression imply several evolutionary mechanisms that, on the one hand, allow for the bias to occur and, on the other hand, condition the dynamics of changes in the affected genes through time [8]. Sex bias in gene expression can be achieved through linkage to sex chromosomes and dosage compensation, sex-specific alternative splicing, and other mechanisms [9–12]. However, these mechanisms primarily affect the expression of regulatory elements, which in turn condition the action of the genes themselves, e.g., following a particular sex-specific splicing or protein maturation pathway. Gene duplication is another mechanism that directly allows for new gene expression profiles, including sex-biased ones [13]. Gene duplication offers an immediate solution to differential expression needs by potentially allowing each copy of a gene to acquire unique functionality. It is now viewed as having played an important, if poorly understood, role in the evolution of sex-biased gene expression [14]. Moreover, gene duplication could also be related, in part, to the relaxation of evolutionary constraints on one of the resulting gene copies, which could, in turn, lead to more rapid gene evolution [15]. Rapid gene evolution has classically been proposed as a consequence of sex-biased and particularly male-biased genes [4,5,16–18]. However, it is not entirely clear whether it is the bias in expression that results in faster evolutionary rates or if it is because of other features of these genes, such as their frequent tissue specificity, which is also correlated with faster evolutionary rates [19].

It is generally accepted that gene duplication is a major force altering the diversity and characteristics of sex-biased genes, but the connection between sex-biased gene expression and evolutionary rates remains poorly understood [8]. So far, these associations have been studied in just a handful of model organisms, and even though it is theoretically plausible that evolutionary processes and functional patterns are related, it is too early to invoke a general rule. Working toward this generalization first requires determining the unequivocal orthology of sex-biased genes between model and non-model species [20]. Furthermore, it requires assuming that orthology and structural homology correlate with functional homology [21,22]. Another problem lies in the actual definition of sex-biased genes. The concept is intuitive and unambiguous: a sex-biased gene is one with different levels of expression between males and females [23]. However, it is also a quantitative one: how different do the expression levels have to be to elicit the activation of the particular evolutionary mechanisms mentioned above? Other non-trivial issues include the occurrence of pleiotropy, the fact that sex-biased genes may be expressed for alternative functions in different tissues and not necessarily related or restricted to one sex, and protein–protein interactions, so that a specific function takes place through physical and genetic modulation by other proteins. Pleiotropy and protein–protein interactions could modulate or limit the evolutionary dynamics of genes, obscuring or changing the expectations derived from the study of model species.

In this study, we aimed to explore the correlation of sex-biased expression with gene duplications and accelerated evolutionary rates in a large evolutionary framework, using non-model organisms for which no gene expression analyses are available. Our work was informed by previous studies involving a model organism (*D. melanogaster*) and used phylogenetic approaches. The obvious candidates for sex-biased genes are those involved in processes that are exclusive to one sex, for example, spermatogenesis in males [18,24]. Thus, in order to test for these differences, we selected a

male reproduction functional group, i.e., a coherent set of genes working together toward a specific reproductive function in males, including genes that are male biased in *Drosophila* spp. and genes that are expressed both in female and in male tissues or non-reproductive tissues. In particular, the present study focused on an integrated male reproductive function—sperm individualization—which is known to involve the action of both constitutive and sex-biased genes in *D. melanogaster* with different degrees of tissue specificity. Sperm individualization is one of the final stages in spermatogenesis that resolves spermatids as individual cells from the syncytial male germline cysts [25]. In a very simplified manner, this process involves a number of stages where (1) a syncytial cyst forms around all spermatids resulting from a primary spermatocyte, (2) an individualization complex formed through cytoskeletal mechanisms and membrane formation encapsulates each of the spermatids, and (3) the syncytial cytoplasm is discarded [26]. We investigated the phylogeny and evolution of these genes across the class Insecta, with particular emphasis on the species-rich order Coleoptera (beetles). The insects we studied included several model organisms for which both orthology assessment and expression studies were publicly available (e.g., modENCODE and OrthoDB projects; [27,28]). Given that beetles are proportionally underrepresented in the genomic and gene profiling literature, we mined relevant data from the 1KITE project (<http://1kite.org/>), thereby broadening representation of beetles in our study and facilitating orthology assessment via phylogenetic approaches [29].

## 2. Materials and Methods

### 2.1. Selection of Functional Group and Expression Profiles

The gene browser AmiGO2 [30] was used to search for genes belonging to the gene ontology category “sperm individualization” (GO:0007291), a category that comprises all genes recognized to participate in the aforementioned processes. With this query, we obtained 54 genes, of which 1 was reported only for mammals (*Spem1*) and was not further considered, and the remaining 53 genes had been previously characterized in *Drosophila melanogaster*. The DNA coding sequences (CDSs) of these genes were retrieved (in September 2017) from FlyBase [31]. A preliminary *blastx* default search was conducted using these CDSs as query sequences, revealing that nine of these genes lacked obvious putative homologs in organisms other than Diptera. These genes (*dj*, *dud*, *fan*, *mst101(3)*, *nkg*, *ntc*, *soti*, *TLL3B*, and *yuri*; named based on *Drosophila* gene nomenclature) were excluded from subsequent analyses. The remaining 44 genes (Table 1) were retained for use in our phylogenetic study and were functionally categorized as (i) unbiased or (ii) sex biased, according to their expression profiles in *Drosophila* using data publicly available in modENCODE [27]. These expression profiles were mined from Affymetrix tiling arrays (Figure 1), designed to study transcription levels in a large number of *Drosophila* cell lines and developmental stages, using modMINE [32]. When the expression profiles of males were less than twofold higher or not more than twofold lower than those measured in females, they were not considered indicative of being biased (a criterion applied in previous studies; e.g., [17]). Five of the genes of interest (*Cul3*, *Dark*, *didum*, *mlt*, and *orb2*) lacked data in the Affymetrix tiling array experiments, and we deduced their sex-based functional profile based on RNA-seq transcriptome profiles available in modENCODE [27].

**Table 1.** Genes belonging to the ontology category “sperm individualization” (GO:0007291) in insects. Genes are identified by their names and their corresponding FlyBase ID in the *Drosophila melanogaster* genome. Information on the general function of the gene and sex biases in expression profiles is also given.

Gene	FlyBase ID	Function	Expression Profile
<i>Act5C</i>	FBgn0000042	cytoskeleton structure	unbiased
<i>Ance</i>	FBgn0012037	peptidase	unbiased
<i>aux</i>	FBgn0037218	ATP binding cofactor of kinase	unbiased
<i>blanks</i>	FBgn0035608	siRNA binding	male biased
<i>Bug22</i>	FBgn0032248	cilium organization and assembly	unbiased

Table 1. Cont.

Gene	FlyBase ID	Function	Expression Profile
<i>CdsA</i>	FBgn0010350	enzyme (CDP diglyceride synthetase)	unbiased
<i>Chc</i>	FBgn0000319	coated vesicles structure	unbiased
<i>ctp</i>	FBgn0011760	dynein complex assembly	unbiased
<i>Cul3</i>	FBgn0261268	protein binding	unbiased
<i>Cyt-c-d</i>	FBgn0086907	electron carrier	male biased
<i>Dark</i>	FBgn0263864	apoptosome assembly	unbiased
<i>didum</i>	FBgn0261397	unconventional myosin	unbiased
<i>Dredd</i>	FBgn0020381	enzyme (caspase)	unbiased
<i>Dronc</i>	FBgn0026404	enzyme (caspase)	unbiased
<i>Duba</i>	FBgn0036180	enzyme (deubiquitinase)	unbiased
<i>EcR</i>	FBgn0000546	transcription factor	unbiased
<i>eIF3m</i>	FBgn0033902	translation initiation factor	unbiased
<i>Fadd</i>	FBgn0038928	protein binding	unbiased
<i>gish</i>	FBgn0250823	enzyme (protein kinase)	unbiased
<i>gudu</i>	FBgn0031905	NA	male biased
<i>heph</i>	FBgn0011224	mRNA binding (translation repression)	unbiased
<i>hmv</i>	FBgn0038607	motile cilium assembly	male biased
<i>jar</i>	FBgn0011225	myosin	unbiased
<i>klhl10</i>	FBgn0040038	substrate recruiting for ubiquitin ligase complex	male biased
<i>Lasp</i>	FBgn0063485	actin/myosin scaffolding	unbiased
<i>Mer</i>	FBgn0086384	cytoskeletal protein binding	unbiased
<i>mlt</i>	FBgn0265512	microtubule removal	unbiased
<i>nes</i>	FBgn0026630	enzyme (lysophospholipid acyltransferase)	unbiased
<i>Npc1a</i>	FBgn0024320	sterol metabolism	unbiased
<i>nsr</i>	FBgn0034740	dynein complex assembly	male biased
<i>orb2</i>	FBgn0264307	translation factor	unbiased
<i>Osbp</i>	FBgn0020626	protein binding	unbiased
<i>oys</i>	FBgn0033476	enzyme (lysophospholipid acyltransferase)	unbiased
<i>Past1</i>	FBgn0016693	membrane assembly	unbiased
<i>Pen</i>	FBgn0011823	protein binding	unbiased
<i>poe</i>	FBgn0011230	calmodulin binding	unbiased
<i>porin</i>	FBgn0004363	membrane channel protein	unbiased
<i>Prosalpha6T</i>	FBgn0032492	enzyme (protease)	male biased
<i>scat</i>	FBgn0011232	protein binding	female biased
<i>shi</i>	FBgn0003392	GTPase for microtubule motility	unbiased
<i>skap</i>	FBgn0037643	ATP binding enzyme	unbiased
<i>sw</i>	FBgn0003654	dynein complex assembly	unbiased
<i>Taz</i>	FBgn0026619	enzyme (phospholipid transacylase)	unbiased
<i>Vps28</i>	FBgn0021814	vesicular trafficking	unbiased

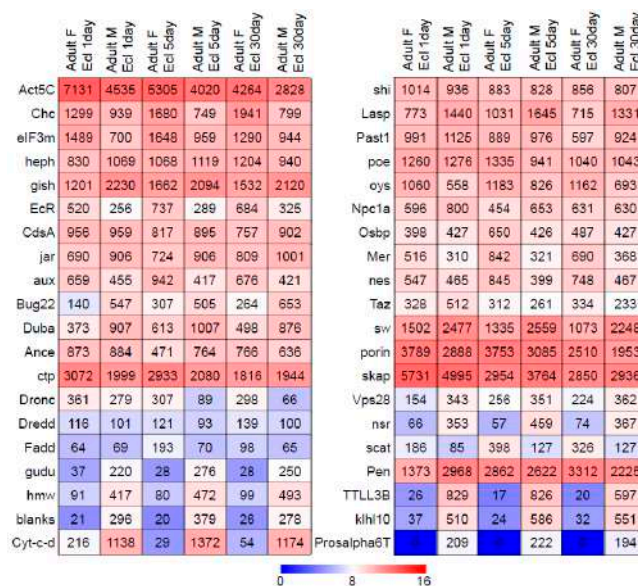


Figure 1. Heatmap visualization of gene expression scores (log<sub>2</sub> of the actual value) of sperm individualization genes in *Drosophila melanogaster* as derived from RNA-seq data from different stages of adult male and female flies [27].



## 2.2. Retrieval of Sperm Individualization Gene Orthologs in Insects

The FlyBase IDs for the 44 genes of interest were used as queries to find putative orthologs and their corresponding eukaryotic orthologous group (EOG) identifiers in OrthoDB v9.1 [33]. We retrieved all insect amino acid sequences for each EOG from the database, together with descriptive information about the number of hits and taxonomic redundancy, as well as data on the relative amino acid sequence divergence of each orthologous group as a proxy for the evolutionary rate in each EOG [28].

The representation of Coleoptera in OrthoDB is currently restricted to six species of three infraorders of the suborder Polyphaga (Table 2). In order to increase the representation of Coleoptera in the sample, we mined the genes of interest from transcriptomic data from beetle species available from 1KITE. The species studied included representatives from all four suborders of Coleoptera (Table 2). Moreover, we also searched for these genes in published RNA-seq data from testis of *Calligrapha multipunctata* (Chrysomelidae), which we expected to be enriched in sperm individualization genes [22]. In order to identify the 44 genes of interest in the assembled beetle transcriptomes, we used the software pipeline Orthograph version 0.5.14 [34]. This software predicts the orthology of nucleotide sequences by mapping their amino acid translation to genes of known ortholog groups using a graph-based best reciprocal hit approach. The pipeline also performs an automatic correction for sequence orientation, frameshifts, and translation. For all Orthograph searches, we used the official gene sets (OGSs) of three reference species: *D. melanogaster* (dmel\_r6.11; <http://flybase.org/>, [35]); the red flour beetle, *Tribolium castaneum* (v3.0; <http://beetlebase.org/>, [36]); and the leaf-cutting ant, *Acromyrmex echinatior* (v3.8; <http://hymenoptera-genome.org/acromyrmex/>, [37,38]).

**Table 2.** Beetle species used in the current study and their current systematic placement. Unless specified otherwise, gene sequence data were obtained from 1KITE.

Suborder Infraorder	Superfamily	Family	Species	Library ID (1KITE)
Archostemata		Micromalthidae	<i>Micromalthus debilis</i>	INSqzbTABRAAPEI-210
Adephaga		Aspidytidae	<i>Sinaspidytes wrasei</i>	WHINSnuyTAAARAPEI-47
		Carabidae	<i>Cicindela hybrida</i>	INShauTBARAPEI-21
		Dytiscidae	<i>Cybister lateralimarginalis</i>	INSnfrTADRAAPEI-16
		Gyrinidae	<i>Gyrinus marinus</i>	INSnfrTBERAAPEI-19
		Noteridae	<i>Noterus clavicornis</i>	INShkeTALRAAPEI-37
Myxophaga		Hydroscaphidae	<i>Hydroscapha redfordi</i>	INSntgTARRAAPEI-208
		Lepiceridae	<i>Lepicerus</i> sp.	INSyvtTAJRAAPEI-19
Polyphaga				
“basal Polyphaga”	Scirtoidea	Scirtidae	<i>Cyphon laevipennis</i>	INSjdsTBDRAAPEI-47
Bostrichiformia	Bostrichoidea	Bostrichidae	<i>Xylobiops basilaris</i>	WHANIsrmTMCLRAAPEI-11
Cucujiformia	Chrysomeloidea	Cerambycidae	<i>Anoplophora glabripennis</i> <sup>a</sup>	-
		Chrysomelidae	<i>Calligrapha multipunctata</i> <sup>b</sup>	-
			<i>Leptinotarsa decemlineata</i> <sup>a</sup>	-
	Cleroida	Byturidae	<i>Byturus ochraceus</i>	INShkeTAORAAPEI-43
		Cleridae	<i>Thanasimus formicarius</i>	INShkeTCERAAPEI-79
	Coccinelloidea	Coccinellidae	<i>Rhyzobius pseudopulcher</i>	WHANIsrmTMABRAAPEI-9
	Curculionoidea	Curculionidae	<i>Dendroctonus ponderosae</i> <sup>a</sup>	-
	Tenebrionoidea	Meloidae	<i>Meloe violaceus</i>	INShauTAYRAAPEI-19
		Tenebrionidae	<i>Tribolium castaneum</i> <sup>a</sup>	-
		Zopheridae	<i>Bitoma cylindrica</i>	WHANIsrmTMAPRAAPEI-39
Elateriformia	Buprestoidea	Buprestidae	<i>Agrilus planipennis</i> <sup>a</sup>	-
	Elateroidea	Lampyridae	<i>Lamprohiza splendidula</i>	INShkeTCGRAAPEI-87
Scarabaeiformia	Scarabaeoidea	Scarabaeidae	<i>Cetonia aurata pisana</i>	WHANIsrmTMAVRAAPEI-53
			<i>Onthophagus taurus</i> <sup>a</sup>	-
Staphyliniformia	Hydrophiloidea	Hydrophilidae	<i>Hydrochara caraboides</i>	INShauTASRAAPEI-13
	Staphylinoidea	Staphylinidae	<i>Ocyopus brunniipes</i>	INShkeTCMRAAPEI-45

<sup>a</sup> Beetle model species and data obtained from OrthoDB; <sup>b</sup> Data available from [22].

Each OGS included the 44 genes belonging to the EOGs of interest. Additionally, Orthograph required a tab-delimited file listing the name of the gene for each EOG and each reference species (obtained from OrthoDB). With this information, Orthograph retrieved from each OGS the genes of

interest and aligned the amino acid sequences to create a profile hidden Markov model with which to conduct a forward search for respective candidate homologs in each of the beetle transcriptomes. The resulting hits were compared with a BLAST search against all genes in all OGSs (reverse search), and for each match between the best hit of the reverse search and the ortholog group of the original forward search, the corresponding transcript was assigned to that specific ortholog group [34]. Each Orthograph search produced the single best hit from each of the 1KITE transcriptomes mined for the study and generated separate files for each EOG, one with the original nucleotide data and one with their amino acid sequence translations, including the sequences of both the beetle targets and the reference species.

### 2.3. Phylogenetic Analyses of Amino Acid Sequences in Insects

Insect amino acid sequences from each EOG and those obtained from the output of Orthograph were aligned with the G-INS-i algorithm of MAFFT v7 [39]. Long autapomorphic insertions in these alignments, possibly corresponding to unrecognized introns, were trimmed manually, as were sequence ends of doubtful quality, typically showing as sequences unaligned beyond one point and longer than the remaining sequences in the alignment, suggesting that the reading frame had been lost and, therefore, the correct start or stop codons were not found either. In a few cases, the protein was retrieved from OrthoDB or the beetle transcripts as disjoint amino acid fragments coming from non-overlapping sequenced transcripts of the same gene. In these cases, the full protein length was reconstituted, and gaps between fragments were filled with missing data. Sequences were secondarily removed from the alignments if they (i) consisted of short fragments usually spanning less than 50% of the gene; (ii) were highly similar and monophyletic for a given species; and/or (iii) were highly divergent in the context of the variability of the alignment, the latter two features assessed based on preliminary phylogenetic analyses of the data.

The resulting purged alignments (deposited in Zenodo.org: 10.5281/zenodo.3380181) were analyzed using SMS [40] to identify the models of amino acid sequence evolution best fitting the data. The resulting models were used in maximum likelihood (ML) tree searches executed using the program PhyML v3.0 [41]. Since some of the genes of interest are multi-copy (in principle, OrthoDB identifies duplicated genes from isoforms resulting from alternative splicing), several gene alignments included many more sequences than taxa, and phylogenetic analyses allowed us to easily recognize when these extra sequences represented gene duplications affecting particular taxa or entire clades. In the former case, one representative of an intraspecific duplication was retained, and in the latter, duplicated versions of the gene were separated into independent alignments, which we realigned with MAFFT. Of the gene variants studied, the one including the sperm individualization gene copy in *Drosophila* was analyzed, assessing the best-fitting evolutionary model again with SMS. ML gene trees were inferred using PhyML, and statistical measures of nodal support were estimated via 100 bootstrap pseudoreplicates.

### 2.4. Phylogenetic Analyses of Nucleotide Sequences in Beetles

Nucleotide sequence matrices of the genes of interest for Coleoptera were generated by combining the sequences retrieved using Orthograph with the corresponding orthologs of model beetle species (Table 2). Data from model beetle species and from a hemipteroid (to be used as an outgroup in the analyses) were obtained with *blastn* searches against the nucleotide collection (nr/nt) at NCBI. The match of the retrieved nucleotide sequences with the amino acid sequence obtained from OrthoDB for the same organisms was confirmed with a subsequent *blastx* search against the reference proteins (refseq\_protein) database, also at NCBI. Nucleotide sequences were aligned using the G-INS-i algorithm implemented in the program MAFFT. Low-quality ends were trimmed and short sequences removed, as above. The aligned sequences were also translated into amino acid sequences to assist the alignment by finding reading frame problems and highly divergent regions, which were secondarily removed.

ML phylogenetic analyses were implemented using these aligned datasets and the same methods described above for the amino acid data.

### 2.5. Estimation of Evolutionary Rates

With very few exceptions, the ML gene trees based on amino acid sequences recovered Hymenoptera and Diptera each as monophyletic and usually with strong (typically 98–100%) bootstrap support. These two clades have particularly well-established age estimates based on independent analyses. They were used as calibration points in Bayesian analyses of evolutionary rates and node dating for each gene tree using the software BEAST v1.8.4 [42]. The nodes for these two clades were consistently constrained as monophyletic in all analyses to avoid uninformative topologies, particularly for genes with low phylogenetic signal, and the calibration densities for the time to their most recent ancestors were modeled as follows. For Hymenoptera, we specified a crown age of 309 Ma (291–347 Ma) after [43], approximately modeled in BEAST as a normal distribution with mean = 309 and Stdev = 10; in turn, the crown age of Diptera was assumed to be 265 Ma (256–269 Ma) according to [44] and approximately modeled as a normal distribution with mean = 265 and Stdev = 5. The analyses used substitution models as determined with SMS, an uncorrelated lognormal relaxed clock [45], and a tree prior under the Yule process. The analyses were run initially for 100 million generations, sampling every 10,000th generation, but in most cases, they had to be replicated and results combined until there was good mixing of parameters and all produced stable estimates with acceptably high effective sample sizes (ESS  $\gg$  200). In a few cases, typically involving datasets that clearly deviated from a molecular clock (i.e., value of `uclid.stdev` > 3), the multiple analyses produced erratic results; here, stable results were obtained using an exponential relaxed distribution. Evolutionary rates, as well as node ages, were calculated using Tracer 1.6 [46] on the annotated maximum clade credibility trees obtained by summarizing the post burn-in trees with LogCombiner 1.8.4 and TreeAnnotator 1.8.4 [42]. Nucleotide substitution rates in beetles were assessed using a similar strategy but with constraining the age of Coleoptera using a normal distribution covering the age range based on the estimate for this order as deduced from the previous analyses. Specifically, we extracted this age as the concordant overlap of all confidence intervals for this parameter in the amino-acid-based trees where Coleoptera was monophyletic.

### 2.6. Statistical Analyses

We tested the hypothesis of no differences in the evolutionary rates of sex-biased genes relative to unbiased genes using a Mann–Whitney  $U$  test [47] at a 0.05 significance level, as implemented in the function “`wilcox.test`” of the R package Stats 3.6.0 [48]. The same test was used to investigate rate differences between genes found as single-copy and as members of multigene families, as well as between genes coordinated in the gene cascade for sperm individualization versus genes participating in this function but not implicated in this interaction network (see below). Finally, genes were tested for differences in absolute evolutionary rates between two main categories based on the overall constancy of those evolutionary rates: genes with relatively homogeneous rates (parameter `uclid.stdev` < 0.6) and genes with heterogeneous rates (`uclid.stdev` > 0.6). These tests were conducted using substitution rates estimated from the insect amino acid data and substitution rates for beetles estimated from nucleotide data. In order to recognize possible interactions of the explanatory variables used in these tests, chi-squared permutation contingency tests of independence were run for each pair of categorical variables used to rank all genes, including expression bias, paralogy, network interaction, and rate heterogeneity. These tests used the “`perm.ind.test`” function of the R package `wPerm` 1.0.1 [49] with 9999 randomization replicates. In all tests, sample sizes allowed for low type I error rates, between 5% and 10% (Power = 0.80).



### 2.7. Analyses Constrained by Gene Interactions

Public databases were used to define the subset of physically or genetically interacting genes among those sharing sperm individualization as a unifying function. Specifically, we established the interaction network of *Drosophila melanogaster* as an interaction model by extracting the information about specific protein–protein physical interactions from BioGRID version 3.4 [50] and that about genetic interactions in metabolic pathways from FlyBase [31]. The obtained graph included 21 nodes (i.e., genes) and 28 edges (i.e., interactions), and the architecture of interactions was used to explore correlations with the evolutionary properties of this subset of genes and with other gene characteristics, including evolutionary rates, patterns of gene duplication, and sex-biased gene expression. Since these genes are not isolated in their function and their interactions, we also considered the total number of known interactions per gene, as shown in BioGRID, as a measure to modulate node importance. We tentatively corrected node importance in every case, calculating the logarithm of the product between node centrality and the absolute number of known interactions per node.

Statistical network analyses were carried out with the aid of R tools implemented in the “igraph” package [51] on the undirected connected graph representing all interacting genes. Measures of node centrality or node “importance” in the networks were obtained relative to the number of receiving edges (“closeness”) or their rank (“eigen\_centrality”). We estimated the correlation between these variables and evolutionary rates and gene paralogy based on the Spearman rank-order correlation coefficients. Additionally, the network community structure was explored with several node modularity optimization algorithms in “igraph”, including the Clauset–Newman–Moore algorithm (command “cluster\_fast\_greedy”) and the Louvain method (command “cluster\_louvain”; [52]), as well as exact modularity maximization (command “cluster\_optimal”) using the algorithm published by [53]. Modularity was also estimated by considering edges instead of nodes and using the algorithm (command “cluster\_edge\_betweenness”) proposed by [54]. We tested for the existence of differences in evolutionary rates for each resulting group using the Kruskal–Wallis test [55]. Additionally, the homogeneity of rates between bipartitions of the network defined by each of the edges separating groups was investigated using a Mann–Whitney *U* test at a 0.05 significance level.

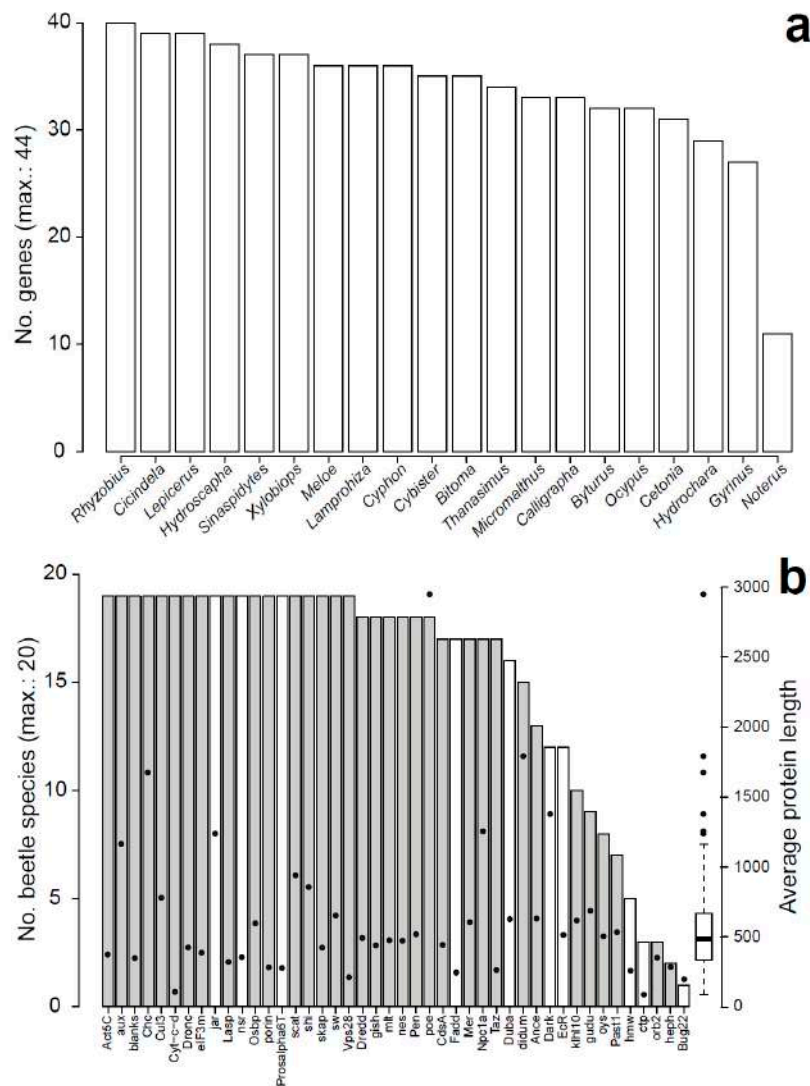
## 3. Results

### 3.1. Characteristics of Datasets: Composition of Sequence Alignments

The 44 investigated genes involved in sperm individualization (GO:0007291) were present in the subclass Pterygota (winged insects), both in Palaeoptera (mayflies and odonates), which were used as outgroups in all analyses, and Neoptera (the remaining orders of winged insects). Most of these genes showed unbiased patterns of gene expression in *Drosophila*, except for eight genes (Table 1). Given the lack of similar functional studies in most other insects, these eight genes represented our hypothesis for biased expression in the insect and beetle datasets. The median length of the associated proteins ranged from 89 amino acids in the case of *ctp* to 2949 amino acids in the case of *poe*, with an average of  $638 \pm 529$  amino acids per protein.

For most of the genes, OrthoDB contributed the amino acid sequences of the six beetle model species to the Coleoptera subset (Figure 2b). The only exceptions were *Bug22*, *ctp*, *Dark*, *EcR*, *Fadd*, *jar*, *nsr*, and *Prosalpha6T*, which lacked data for one of the species, *Duba* for two, and *hmv* for five. Mapping of orthologous genes using Orthograph from transcriptomes of a selection of 19 beetle species from the 1KITE Project and one testis-specific transcriptome from another beetle species resulted in positive hits in all cases, although with different success rates, possibly related to the quality or source of the transcriptomes. No single species yielded ortholog sequence data for all tested genes, with *Rhyzobius pseudopulcher* retrieving the highest number of genes (40 out of 44) and two water beetle species, *Gyrinus marinus* and *Noterus clavicornis*, retrieving the lowest (27 and 11 genes, respectively). For 70% of the beetle species, we retrieved at least 75% of the genes (Figure 2a). In turn, for all genes analyzed, we found orthologs in the beetle transcriptomes, but with different success rates (Figure 2b).

A large proportion of genes (43.2%) were found in at least 19 out of 20 beetle species, and most of them (72.7%) were found in 15 or more beetle species. Conversely, eight genes could not be found in at least half of the species analyzed, with genes such as *hmv*, *ctp*, *orb2*, *heph*, and *Bug22*, showing the lowest recovery frequencies ( $n \leq 5$ ). The proteins encoded by these genes were shorter than the average but were also typically lacking recognized orthologs in some of the beetle model species (Figure 2b).



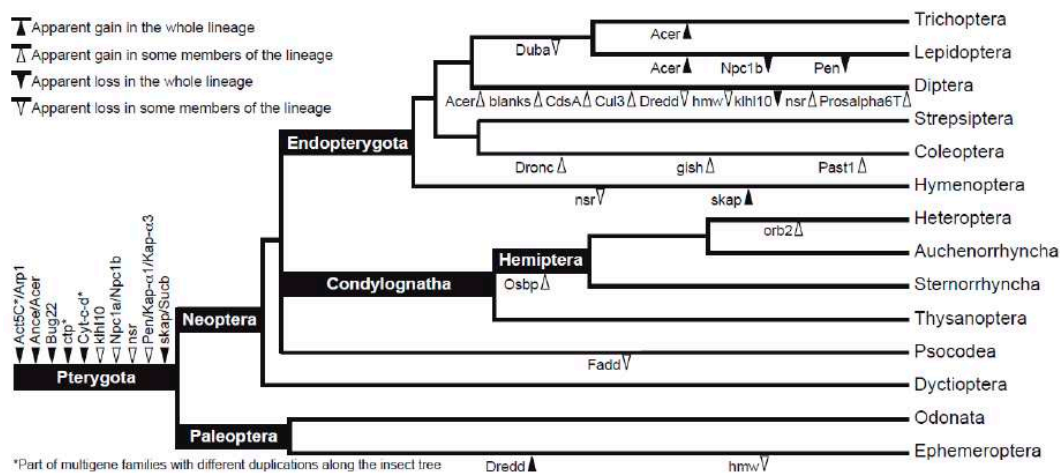
**Figure 2.** Performance of Orthograph searches of sperm individualization gene orthologs in beetle transcriptomes. (a) Number of genes retrieved from each of the non-model beetle species transcriptomes. (b) Number of beetle species yielding ortholog sequences for each of the sperm individualization genes analyzed, with information on the average protein length and showing genes absent in one or more OrthoDB beetle model species as white columns.

### 3.2. Characteristics of Datasets: Gene Duplications

At the time of this study, OrthoDB curated data for 119 insect species. When more than 119 sequences were retrieved for a particular gene, this informed in most cases of potential multi-copy genes (Table 3). Their actual presence was confirmed in the ML trees when including all the sequences retrieved from OrthoDB and the beetle sequences mined from 1KITE. In these cases, we used the annotation of the *D. melanogaster* sequence to recognize the sperm individualization paralog of interest. Figure 3 shows a diagram of gene duplications (and some secondary gene losses) as recognized in this study.

**Table 3.** Summary of sequence characteristics of genes retrieved from OrthoDB. The table lists the number of sequences (N) and species (Sp; with a maximum of 119 species), number of species in which the gene is single copy (Single), the median protein length (L), and relative evolutionary rate (r) as tabulated in OrthoDB. Furthermore, the number (n) of aligned sequences in this study and the alignment lengths (Length), as well as the inferred optimal evolutionary model, are given.

Gene	N	Sp	Single	L	r	n	Length	Model
<i>Act5C</i>	517	115	7	376	0.60	-	-	-
<i>Ance</i>	238	109	30	631	0.92	90	621	LG + G + I
<i>aux</i>	136	112	94	1164	1.21	126	1755	JTT + G + I + F
<i>blanks</i>	172	107	71	348	1.55	122	719	JTT + G + I + F
<i>Bug22</i>	201	115	36	200	0.61	83	270	LG + G + I
<i>CdsA</i>	118	109	101	445	0.76	122	574	JTT + G + I + F
<i>Chc</i>	123	115	110	1676	0.63	130	1726	JTT + G + I
<i>ctp</i>	121	98	79	89	0.57	-	-	-
<i>Cul3</i>	153	116	91	780	0.73	133	858	JTT + G + I
<i>Cyt-c-d</i>	148	110	74	108	0.65	-	-	-
<i>Dark</i>	119	106	94	1378	1.99	112	2193	JTT + G + I + F
<i>didum</i>	126	113	102	1793	1.10	124	2175	LG + G + I
<i>Dredd</i>	91	81	75	493	1.76	98	676	JTT + G + I + F
<i>Dronc</i>	128	96	78	425	1.67	119	654	WAG + G + I + F
<i>Duba</i>	105	99	93	628	0.95	113	1087	JTT + G + I + F
<i>EcR</i>	121	113	105	515	0.83	120	576	JTT + G + I
<i>eIF3m</i>	116	113	111	387	0.77	130	394	JTT + G + I
<i>Fadd</i>	96	93	90	246	1.77	108	353	JTT + G + I + F
<i>gish</i>	129	113	99	441	0.73	124	401	JTT + G + I
<i>gudu</i>	125	115	106	689	1.01	124	658	LG + G + I
<i>heph</i>	206	114	47	285	0.78	114	597	JTT + G + I + F
<i>hmw</i>	78	76	74	260	1.38	80	925	JTT + G + I + F
<i>jar</i>	131	111	97	1238	0.86	129	1375	JTT + G + I + F
<i>klhl10</i>	214	109	54	619	0.96	113	630	LG + G + I
<i>Lasp</i>	113	105	97	321	0.79	121	298	JTT + G + I
<i>Mer</i>	120	112	105	605	0.87	129	686	JTT + G + I
<i>mlt</i>	124	112	103	477	1.10	130	651	LG + G + I + F
<i>nes</i>	122	112	104	474	1.21	128	472	LG + G + I + F
<i>Npc1a</i>	216	116	23	1256	0.98	124	1435	LG + G + I
<i>nsr</i>	317	115	38	355	0.92	129	563	JTT + G + I
<i>orb2</i>	115	106	98	351	0.62	105	293	JTT + G + I
<i>Osbp</i>	158	112	79	597	0.91	130	1094	JTT + G + I
<i>oys</i>	121	108	97	505	1.03	115	463	LG + G + I
<i>Past1</i>	124	114	104	534	0.66	120	564	LG + G + I
<i>Pen</i>	342	116	7	519	0.83	121	593	LG + G + I + F
<i>poe</i>	183	115	84	2949	1.08	129	3846	JTT + G + I + F
<i>porin</i>	124	108	98	282	0.86	127	286	LG + G + I + F
<i>Prosalpha6T</i>	126	108	91	277	0.77	125	312	LG + G + I + F
<i>scat</i>	130	116	103	942	1.11	133	1233	JTT + G + I
<i>shi</i>	133	113	94	857	0.67	132	1005	LG + G + I
<i>skap</i>	247	116	8	424	0.80	128	476	LG + G + I
<i>sw</i>	125	115	109	655	0.80	133	755	JTT + G + I
<i>Taz</i>	105	102	99	265	0.91	118	303	LG + G + I + F
<i>Vps28</i>	135	112	94	212	0.72	131	213	LG + G + I



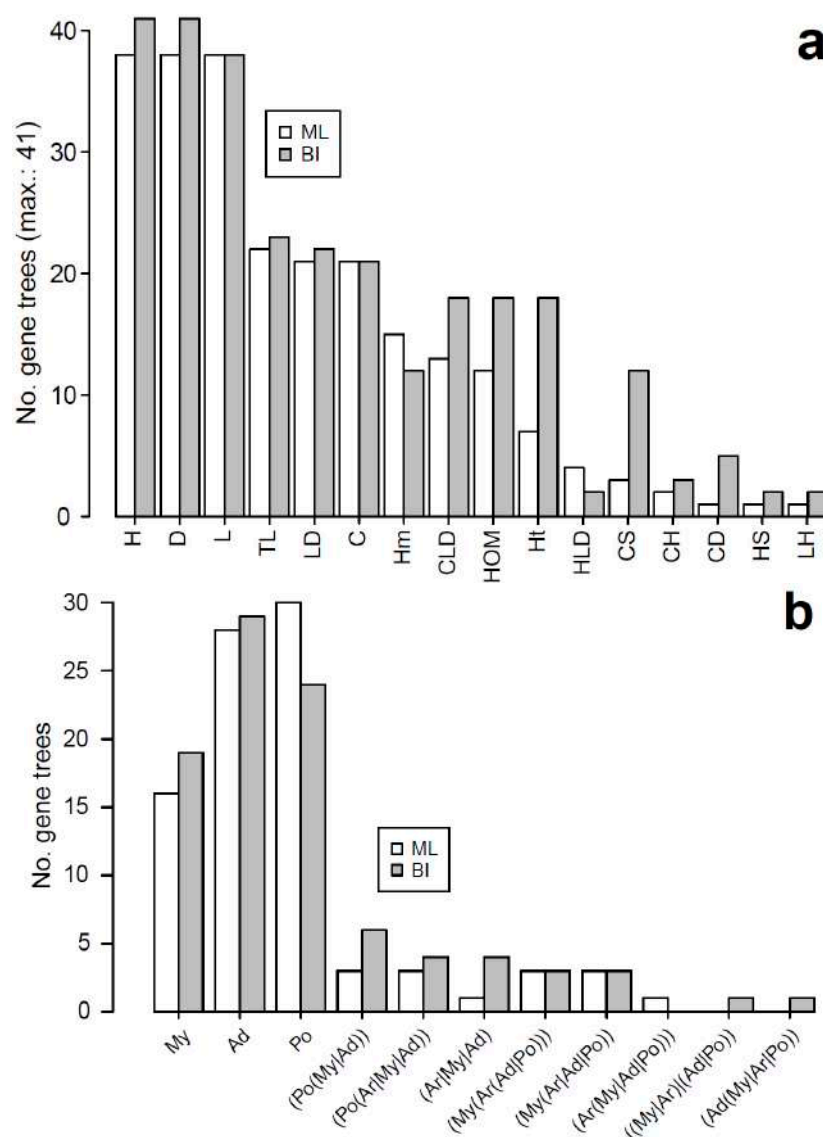
**Figure 3.** Schematic consensus phylogeny of insects (drawn from trees in [56–58]) showing the inferred evolutionary gains and losses of sperm individualization genes for major insect lineages.

For a total of 21 genes, we had no evidence for duplications or losses in the winged insect lineage. However, 18 genes showed duplications in Pterygota or parts of this evolutionary lineage. Two of these genes, *Act5C* and *Cyt-c-d*, were found as part of multigene families, and it was difficult to tell individual copies apart with the available data and as a result of their high similarity. *Act5C* is part of a gene complex with a deep split in all insects, including actin-related proteins (Arp1 in *Drosophila*) and several actins resulting from various duplications. We found evidence for at least six actin-like gene copies in Acalyptratae flies (fruit and peacock flies, among others), five in mosquitoes, four in Hymenoptera and Palaeoptera, at least three in Coleoptera and the hemipteroids, and at least two among Lepidoptera. *Act5C*, in particular, is highly conserved in the whole of Pterygota, and most of the available beetle sequences were retrieved close to this specific *Drosophila* paralog in the phylogeny. In turn, *Cyt-c-d* was revealed as a member of a multigene family in most insect groups, including odonates, some hemipteroids, beetles, and some dipterans. The proteins encoded by these genes are short and highly conserved, so paralogs could not be resolved easily, but most beetle sequences retrieved by Orthograph were more similar to the *Cyt-c-p* copy of the gene in *Drosophila*. Finally, *ctp* corresponded to a very short fragment, highly conserved and with evidence for paralogy, though it was not possible to discriminate gene copies. Given the difficulty of discerning orthologs, these three genes were not considered in downstream analyses.

Seven of the duplicated genes—namely, *Ance/Acer*, *Bug22*, *klhl10*, *Npc1a/Npc1b*, *nsr*, *Pen/Kap-α1/Kap-α3*, and *skap/Sucb*—were duplicated in all studied insects and, in some cases, with one of the copies being subsequently lost or further multiplied in particular lineages. For example, orthologs of the *Npc1b* and *Pen* copies were lost in Lepidoptera, the sister copy of *klhl10* was lost in Diptera, and one copy of *nsr* was lost in aculeate Hymenoptera. Acalyptratae (Diptera) had three additional copies of *nsr* (four in *Bactrocera* tephritid peacock flies); *Acer* was duplicated independently in Trichoptera, Lepidoptera, and some Diptera; and *skap* had an additional copy among the Hymenoptera. Overall, 10 genes had lineage-specific duplications. *Dredd* had several copies in *Ephemera* alone; *orb2* and *Osbp* were duplicated in some hemipterans; and for *Dronc*, *gish*, and *Past1*, we found evidence for duplications in Coleoptera. Finally, the remaining four genes were duplicated in Diptera: *blanks* and *Cul3*, with fast-evolving copies in some dipterans; *CdsA* in some nematocerans (midges and moth-flies); and *Prosalpha6* in *Drosophila* alone (wherein only the paralog *Prosalpha6T*, perhaps missing in all the other insects, is male biased). Apart from the lineage-specific losses found for *Npc1b*, *Pen*, and the sister copies of *klhl10* and *nsr*, other gene losses detected in our data set affected *Dredd* (missing in mosquitoes [Diptera: Culicidae]), *Duba* (lacking in Trichoptera and Lepidoptera), *Fadd* (absent in some Hemiptera), and *hmw* (not recorded in *Ephemera* [Ephemeroptera] or *Anopheles* [Diptera]).

### 3.3. Evolutionary Rates of Sperm Individualization Genes in Insecta and Coleoptera

Amino acid sequence matrices of the orthologous sperm individualization genes of insects and nucleotide sequence data of Coleoptera were used to infer gene trees under ML and Bayesian inference and to estimate evolutionary rates (Supplementary Files S1–S4). In general, both methods produced similar gene trees, e.g., with respect to resolving the relationships of the insect orders and some infraordinal relationships (Figure 4a), usually with relatively strong nodal support, and consistent with the current systematic knowledge for insects [56]. However, most trees had relatively poorly resolved deep relationships, particularly within the hemimetabolous insect orders, which were represented by relatively few taxa. In turn, in most beetle trees, the suborders represented by several species were retrieved as monophyletic, but there was no consensus among trees on subordinal relationships (Figure 4b). However, in most cases the topologies were consistent with Polyphaga being sister to the other three suborders (Adephaga, Myxophaga, and Archostemata).



**Figure 4.** Frequency of maximum likelihood (ML) and Bayesian inference (BI) sperm individualization gene trees resolving particular higher taxa or relationships among them in insects (a) and beetles (b). Key: Ad, Adephaga; Ar, Archostemata; C, Coleoptera; D, Diptera; H, Hymenoptera; Hm, Hemiptera; HOM, Holometabola; Ht, Heteroptera; L, Lepidoptera; My, Myxophaga; Po, Polyphaga; S, Strepsiptera; T, Trichoptera.

Based on the previous phylogenies, the amino acid substitution rates for 41 proteins encoded by sperm individualization genes (excluding the 3 proteins for which orthology could not be confirmed) spanned nearly two orders of magnitude, from 0.000237 amino acid changes per lineage and million years (subs./l./Ma) in the protein orb2 to 0.009667 subs./l./Ma in the protein hmw (Table 4). The average substitution rate for the whole dataset was  $0.00239 \pm 0.003012$  subs./l./Ma. Slightly over half (56%) of these proteins, typically those with lower overall substitution rates, exhibited evolutionary rates inconsistent with a molecular clock, i.e., rates on individual branches with more substantial departures from the mean ( $uclid.stdev \geq 0.6$ ).

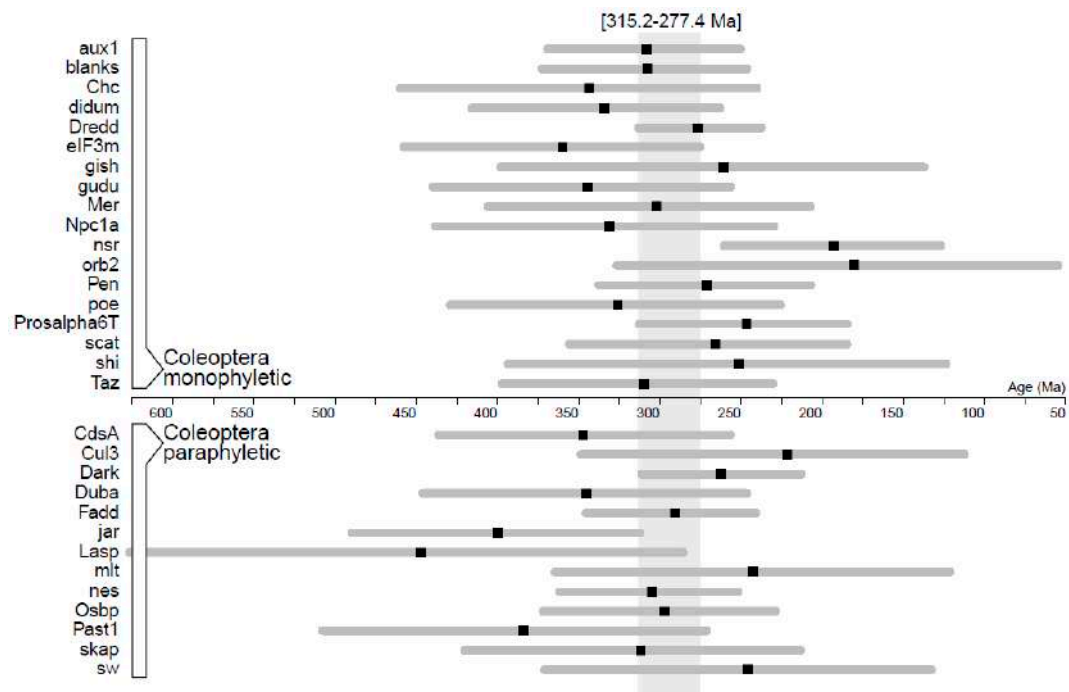
**Table 4.** Characteristics of amino acid datasets of sperm individualization proteins of insects deduced from information in public databases (B: unbiased [0] and sex-biased [1] genes; N: non-interacting [0] and interacting [1]), as well as information deduced from their phylogenetic analyses, including duplications (D: single [0] and multicopy [1]), evolutionary rates, evolutionary rate heterogeneity ( $uclid.stdev$ ), and the estimated age of the clade Coleoptera.

Gene	B/D/N	Substitution Rate ( $\times 10^{-3}$ )	$uclid.stdev$	Age Coleoptera
<i>hmw</i>	1/0/0	$9.67 \pm 1.349$	2.972	-
<i>Dark</i>	0/0/1	$5.27 \pm 0.282$	0.418	264.2 [214.3–314.2]
<i>blanks</i>	1/1/1	$4.38 \pm 0.382$	0.483	310.6 [248.2–376.1]
<i>Dredd</i>	0/0/1	$4.23 \pm 0.219$	0.309	278.4 [239.2–315.8]
<i>Fadd</i>	0/0/1	$4.21 \pm 0.277$	0.345	-
<i>Dronc</i>	0/1/1	$3.62 \pm 0.221$	0.391	379.8 [339.5–426.0] <sup>b</sup>
<i>Duba</i>	0/0/1	$3.18 \pm 0.257$	0.566	348.2 [248.5–450.2] <sup>b</sup>
<i>nsr</i>	1/1/0	$3.00 \pm 0.279$	0.672	194.1 [127.4–262.5]
<i>Bug22</i>	0/1/0	$2.82 \pm 0.278$	0.559	-
<i>scat</i>	1/0/0	$2.73 \pm 0.259$	>3 <sup>a</sup>	267.8 [185.8–359.0]
<i>aux</i>	0/0/1	$2.45 \pm 0.154$	0.386	311.0 [252.3–372.7]
<i>nes</i>	0/0/0	$2.27 \pm 0.130$	0.423	307.2 [253.7–365.2] <sup>b</sup>
<i>poe</i>	0/0/0	$2.10 \pm 0.174$	>3 <sup>a</sup>	328.7 [227.3–433.1]
<i>Osbp</i>	0/1/0	$2.05 \pm 0.180$	0.601	299.6 [230.3–375.4] <sup>b</sup>
<i>Npc1a</i>	0/1/0	$1.91 \pm 0.157$	>3 <sup>a</sup>	334.2 [231.5–442.6]
<i>didum</i>	0/0/1	$1.90 \pm 0.126$	0.511	337.0 [264.9–419.9]
<i>Pen</i>	0/1/0	$1.78 \pm 0.120$	0.531	273.2 [208.2–340.5]
<i>klhl10</i>	1/1/1	$1.69 \pm 0.124$	0.816	-
<i>Prosalpha6T</i>	1/1/0	$1.68 \pm 0.151$	0.559	248.5 [185.6–315.1]
<i>oys</i>	0/0/0	$1.57 \pm 0.127$	0.558	-
<i>gudu</i>	1/0/0	$1.45 \pm 0.115$	0.553	347.7 [258.1–444.1]
<i>Ance</i>	0/1/0	$1.42 \pm 0.099$	0.396	-
<i>sw</i>	0/0/1	$1.38 \pm 0.171$	3.712	247.5 [133.2–374.3]
<i>Taz</i>	0/0/0	$1.33 \pm 0.111$	0.607	312.2 [231.9–401.0]
<i>Mer</i>	0/0/1	$1.30 \pm 0.130$	0.936	304.5 [208.4–409.9]
<i>skap</i>	0/1/1	$1.15 \pm 0.116$	0.557	314.4 [214.9–424.3] <sup>b</sup>
<i>CdsA</i>	0/0/0	$1.15 \pm 0.103$	0.629	350.5 [258.6–440.6] <sup>b</sup>
<i>jar</i>	0/0/1	$1.06 \pm 0.086$	0.482	403.2 [315.0–494.6] <sup>b</sup>
<i>porin</i>	0/0/0	$0.99 \pm 0.106$	0.779	-
<i>Lasp</i>	0/0/1	$0.98 \pm 0.138$	0.930	451.5 [288.4–633.0] <sup>b</sup>
<i>Cul3</i>	0/1/1	$0.98 \pm 0.130$	3.798	223.1 [112.6–351.7] <sup>b</sup>
<i>EcR</i>	0/0/0	$0.93 \pm 0.091$	0.789	-
<i>heph</i>	0/0/0	$0.90 \pm 0.118$	3.919	-
<i>eIF3m</i>	0/0/1	$0.86 \pm 0.084$	0.480	363.0 [277.4–462.1]
<i>shi</i>	0/0/1	$0.74 \pm 0.092$	3.847	253.2 [124.3–397.0]
<i>Past1</i>	0/1/1	$0.71 \pm 0.069$	0.713	387.4 [273.0–513.2] <sup>b</sup>
<i>Vps28</i>	0/0/0	$0.63 \pm 0.079$	0.822	-
<i>gish</i>	0/1/0	$0.47 \pm 0.070$	4.174	262.9 [137.6–401.9]
<i>mlt</i>	0/0/0	$0.42 \pm 0.523$	3.268	244.6 [121.4–368.1] <sup>b</sup>
<i>Chc</i>	0/0/1	$0.32 \pm 0.035$	0.700	346.6 [242.1–464.4]
<i>orb2</i>	0/1/0	$0.24 \pm 0.035$	1.414	180.4 [52.1–328.6]

<sup>a</sup> Data analyzed under exponential relaxed clock, with  $uclid.stdev$  estimated from inconclusive runs under an uncorrelated lognormal relaxed clock; <sup>b</sup> Coleoptera is rendered paraphyletic by the inclusion of Strepsiptera.



The analyses of evolutionary rates yielded age estimates for the clade Coleoptera with averages ranging between 180.4 Ma, in the case of *orb2*, and 451.5 Ma, in the case of *Lasp*, with broad confidence intervals of  $186.2 \pm 62.49$  Ma on average (Table 4). Coleoptera was recovered as monophyletic in 18 of the analyses, and the overlap of the age confidence intervals obtained for each gene covered a period between 277.4 and 315.2 Ma (except in the case of *nsr*, which yielded an age much younger than the oldest known beetle fossils) (Figure 5). This time interval was used to restrict the age of Coleoptera in subsequent analyses, and it was consistent with most clade age estimates for Coleoptera obtained in analyses where the beetle clade also included Strepsiptera.



**Figure 5.** Inferred ages and 95% credibility intervals for the Coleoptera clade (top panel) or a Coleoptera + Strepsiptera clade (bottom panel) based on the molecular clock analyses of amino acid sequence data of sperm individualization genes. The full overlap of age estimates of monophyletic Coleoptera identifies an interval (shaded area) consistent with the proposed age of the group based on fossil data and used here as age prior for the evolutionary rate analyses in beetles.

The above time constraint for Coleoptera produced instantaneous nucleotide substitution rates ranging from 0.00208 subs./l./Ma in the case of the gene *nes* to 0.01190 subs./l./Ma in the case of *Cul3*, with an average substitution rate for the whole set of genes investigated of  $0.00452 \pm 0.002083$  subs./l./Ma (Table 5). Slightly over half these genes had substitution rates relatively consistent with a molecular clock ( $uclid.stdev < 0.6$ ), and in contrast to the case of the amino acid sequence analyses, the genes departing from the molecular clock were those with higher nucleotide substitution rates.

**Table 5.** Characteristics of the nucleotide phylogenetic data sets of sperm individualization genes in Coleoptera. The number of species (N), length of nucleotide sequence alignments (L), the determined evolutionary model, inferred evolutionary rates, and information on rate heterogeneity (ucl.d.stdev) are given for each gene.

Gene	N	L	Model	Substitution Rate ( $\times 10^{-3}$ )	ucl.d.stdev
<i>Cul3</i>	25	2196	TN93 + G + I	11.90 $\pm$ 2.603	2.707
<i>gish</i>	24	1197	GTR + G + I	9.34 $\pm$ 1.981	2.821
<i>Act5C</i>	15	1128	GTR + G + I	8.91 $\pm$ 2.232	2.839
<i>scat</i>	25	1974	GTR + G + I	7.26 $\pm$ 1.397	2.958
<i>eIF3m</i>	25	1155	GTR + G + I	7.06 $\pm$ 1.394	2.872
<i>poe</i>	23	3291	GTR + G + I	6.90 $\pm$ 1.322	2.844
<i>Dredd</i>	23	732	GTR + G + I	6.41 $\pm$ 1.255	3.012
<i>CdsA</i>	21	1323	GTR + G + I	6.35 $\pm$ 1.229	2.954
<i>blanks</i>	25	672	GTR + G + I	5.82 $\pm$ 1.172	2.994
<i>shi</i>	25	2610	GTR + G + I	5.81 $\pm$ 1.161	2.918
<i>skap</i>	24	1302	GTR + G + I	5.09 $\pm$ 0.984	3.033
<i>Dark</i>	14	1863	GTR + G + I	4.96 $\pm$ 0.905	2.947
<i>Duba</i>	20	858	GTR + G + I	4.93 $\pm$ 0.997	2.916
<i>Prosalpha6T</i>	25	822	GTR + G + I	4.90 $\pm$ 0.959	3.041
<i>Chc</i>	22	4944	GTR + G + I	4.78 $\pm$ 0.410	0.270
<i>klhl10</i>	14	1779	GTR + G + I	4.48 $\pm$ 0.849	2.952
<i>jar</i>	25	3393	GTR + G + I	4.47 $\pm$ 0.840	3.005
<i>didum</i>	21	5073	GTR + G + I	4.41 $\pm$ 0.817	3.029
<i>oys</i>	14	1347	GTR + G + I	4.41 $\pm$ 0.839	3.019
<i>ctp</i>	8	267	GTR + G	3.79 $\pm$ 1.107	0.232
<i>mlt</i>	24	1248	GTR + G + I	3.61 $\pm$ 0.464	0.474
<i>sw</i>	25	1455	GTR + G + I	3.61 $\pm$ 0.347	0.344
<i>Taz</i>	23	774	GTR + G + I	3.52 $\pm$ 0.278	0.340
<i>Fadd</i>	19	228	GTR + G + I	3.41 $\pm$ 0.495	0.463
<i>nsr</i>	24	780	GTR + G + I	3.39 $\pm$ 0.464	0.409
<i>orb2</i>	8	834	GTR + G + I	3.36 $\pm$ 0.834	0.477
<i>Npc1a</i>	23	3756	GTR + G + I	3.32 $\pm$ 0.280	0.255
<i>Dronc</i>	24	954	GTR + G + I	3.17 $\pm$ 0.265	0.221
<i>EcR</i>	17	1278	GTR + G + I	3.13 $\pm$ 0.396	0.382
<i>Vps28</i>	26	573	GTR + G + I	3.12 $\pm$ 0.509	0.169
<i>Osbp</i>	25	1881	GTR + G + I	3.09 $\pm$ 0.285	0.365
<i>Past1</i>	13	1566	GTR + G + I	3.09 $\pm$ 0.344	0.186
<i>aux</i>	18	2130	GTR + G + I	3.09 $\pm$ 0.295	0.252
<i>Lasp</i>	25	423	GTR + G + I	3.08 $\pm$ 0.495	0.110
<i>gudu</i>	15	1848	GTR + G + I	2.93 $\pm$ 0.280	0.450
<i>Pen</i>	24	1440	GTR + G + I	2.92 $\pm$ 0.252	0.296
<i>porin</i>	25	849	GTR + G + I	2.78 $\pm$ 0.393	0.499
<i>Mer</i>	23	1701	GTR + G + I	2.78 $\pm$ 0.260	0.329
<i>Ance</i>	18	1716	GTR + G + I	2.59 $\pm$ 0.198	0.299
<i>hmw</i>	8	201	GTR + G	2.57 $\pm$ 0.457	0.118
<i>nes</i>	25	1317	GTR + G + I	2.08 $\pm$ 0.171	0.538

### 3.4. Analysis of Rate Differences

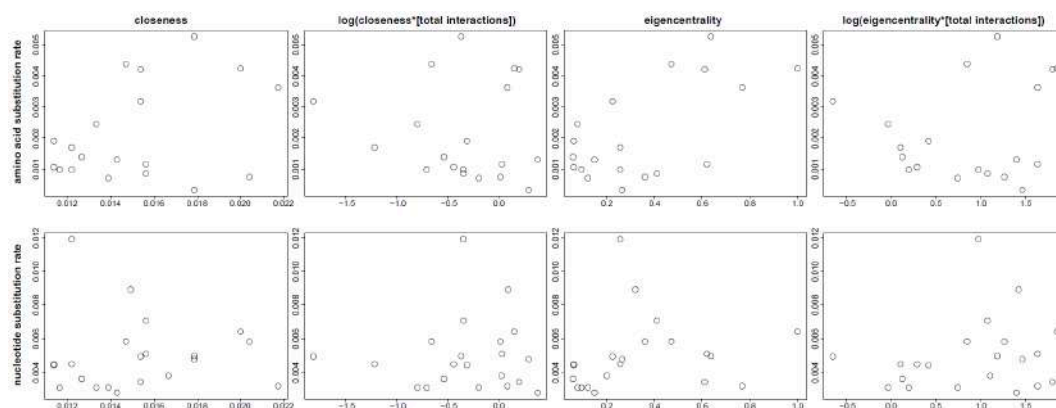
Permutation tests of independence produced non-significant results for every pair of independent variables used in subsequent tests, suggesting that there were no interactions among them. The null hypothesis that sperm individualization genes with or without duplications in the insect lineage had the same evolutionary rates was not rejected (Mann–Whitney  $U = 183$ ,  $p = 0.758$ ; also treating hemipteroid *orb2* and *Osbp* duplications as non-duplicated genes:  $U = 157$ ,  $p = 0.497$ ). Similarly, this hypothesis was not rejected in the case of genes working in coordination in a gene interaction network (like the one deduced for *Drosophila*) tested against genes dissociated from this network (Mann–Whitney  $U = 190$ ,  $p = 0.632$ ). However, when genes were split into two categories according to their predicted



sex expression bias, or according to whether they evolved in a clocklike fashion, the null hypothesis of no differences in their evolutionary rates was rejected at the 0.05 significance level (Mann–Whitney  $U = 52$ ,  $p = 0.019$  and Mann–Whitney  $U = 323$ ,  $p = 0.002$ , respectively). In these cases, sex-biased and clock-constrained genes would have slightly faster rates, except for the male-biased gene *hmv*, a fast-evolving protein departing nonetheless from a molecular clock. The same tests, when applied to nucleotide substitution rates of the genes of interest in beetles, produced non-significant results when rate differences were tested for predicted expression biases ( $U = 115$ ,  $p = 0.817$ ), gene duplications in the beetle lineage ( $U = 181$ ,  $p = 0.551$ ), or their predicted coordination in an interaction network ( $U = 156$ ,  $p = 0.118$ ). The test produced a clear significant result when rate differences were tested against the clocklike behavior of data ( $U = 5$ ,  $p < 0.001$ ), with the genes departing from the molecular clock having much higher rates (genes[ucl.d.stdev < 0.6]:  $0.00314 \pm 0.000533$  versus genes[ucl.d.stdev  $\geq 0.6$ ]:  $0.00620 \pm 0.002030$ ).

### 3.5. Evolutionary Patterns in the Sperm Individualization Interaction Network

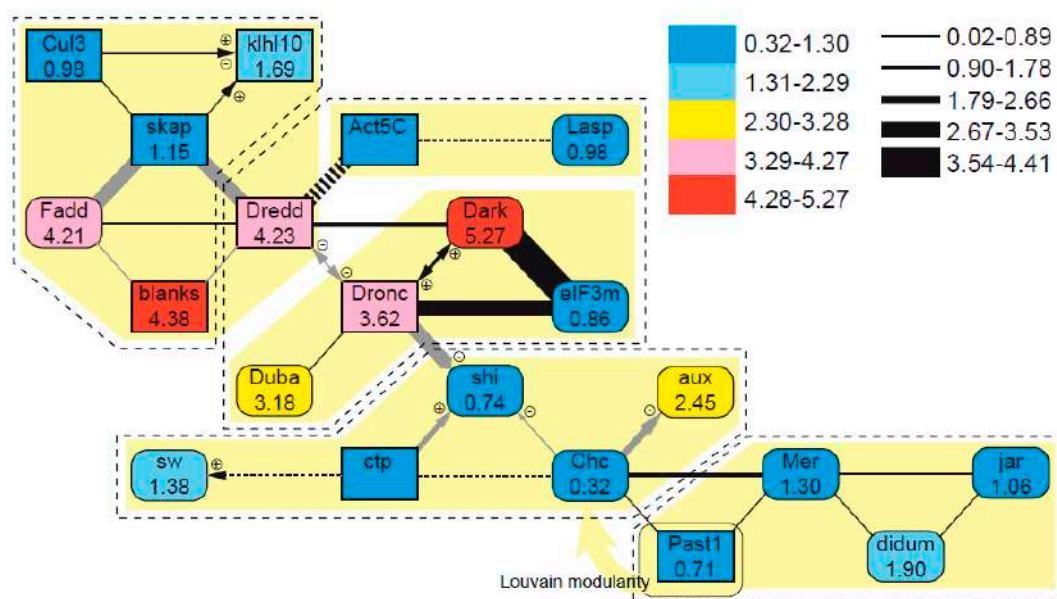
The results of the Spearman's rank correlation tests between amino acid substitution rates and measures of node importance based on the number of receiving edges ( $S = 956.6$ ,  $\rho = 0.1609$ ,  $p = 0.5106$ ), rank ( $S = 712.6$ ,  $\rho = 0.3749$ ,  $p = 0.1138$ ), or their respective corrections considering the total number of genetic interactions of the nodes of interest ( $S[\text{edges}] = 1340.0$ ,  $\rho = -0.1754$ ,  $p = 0.4709$ ;  $S[\text{rank}] = 1054.0$ ,  $\rho = 0.0754$ ,  $p = 0.7592$ ) were all non-significant (Figure 6). Similarly, the correlations between nucleotide substitution rates in beetles and node centrality measures based on the number of receiving edges ( $S = 1115.3$ ,  $\rho = 0.2758$ ,  $p = 0.2263$ ) or their tentative correction based on edges ( $S = 1520$ ,  $\rho = 0.0130$ ,  $p = 0.9573$ ) and rank ( $S = 1180$ ,  $\rho = 0.2338$ ,  $p = 0.3063$ ) were non-significant. However, when node centrality was assessed based on the first eigenvector of the adjacency matrix, their correlation with nucleotide substitution rates was significant ( $S = 770.5$ ,  $\rho = 0.4997$ ,  $p = 0.0211$ ), suggesting a slight effect of more densely connected regions of the network having significantly higher evolutionary rates. In turn, there was no evidence for a correlation between genes being single copy or duplicated and any centrality measure without (edges:  $S = 1772.9$ ,  $\rho = -0.1512$ ,  $p = 0.5129$ ; rank:  $S = 2127.7$ ,  $\rho = -0.3816$ ,  $p = 0.0878$ ) or with correction (edges:  $S = 1368.7$ ,  $\rho = 0.1112$ ,  $p = 0.6312$ ; rank:  $S = 1980.5$ ,  $\rho = -0.2860$ ,  $p = 0.2088$ ).



**Figure 6.** Biplots of the correlation between different measures of node importance in the interaction network of sperm individualization genes and their amino acid evolutionary rates in insects (top panels) and nucleotide substitution rates in beetles (bottom panels).

Network modularity measures split the gene interaction network into four groups when using edge-based partitioning or five groups when using node-based partitioning, with considerable agreement between strategies (Figure 7). Exact modularity and the Clauset–Newman–Moore algorithm produced identical groupings, differing from the edge-based solution in the transfer of one node (*Dredd*) to an adjacent group and the split of two nodes (*Act5C* and *Lasp*) as an additional group. The

Louvain modularity produced groups identical to the other node-based methods, but transferring one node (*Past1*) into the adjacent group. None of these global partitioning strategies showed statistical differences in amino acid substitution rates (“edge\_betweenness”, chi-sq = 2.8835,  $p = 0.4099$ ; “optimal”, chi-sq = 2.5958,  $p = 0.6276$ ; “louvain”, chi-sq = 2.4258,  $p = 0.6580$ ) or nucleotide substitution rates (“edge\_betweenness”, chi-sq = 4.9143,  $p = 0.1782$ ; “optimal”, chi-sq = 5.4316,  $p = 0.2458$ ; “louvain”, chi-sq = 4.9848,  $p = 0.2889$ ). However, when different group bipartitions of the network were considered, the edge between *Dronc* and *shi* (ABC and DE clusters in Figure 7) delimited groups with different amino acid substitution rates ( $U = 72$ ,  $p = 0.0279$ ) and different nucleotide substitution rates when beetle data were considered both for edge ( $U = 83$ ,  $p = 0.0409$ ) and for node ( $U = 93$ ,  $p = 0.0062$ ) partitions. Nucleotide substitution rates were also statistically significantly different across the edges joining *Dredd* and *Dronc* ( $U = 80$ ,  $p = 0.0200$ ; AB and CDE clusters in Figure 7) and *Chc* and *Past1* ( $U = 69$ ,  $p = 0.0147$ ; ABCD and E clusters in Figure 7).



**Figure 7.** Mutual interaction network of sperm individualization genes in *Drosophila* including protein–protein (lines) and genetic/regulatory interactions (arrows), the latter with information on the enhancing and/or repressing modulation effects. Nodes represent interacting proteins, and they are color coded according to their inferred amino acid evolutionary rates. Edges represent documented interactions between proteins, with their width being proportional to the evolutionary rate differences between interacting proteins (dashed lines are used when their evolutionary rate data are missing). Dashed-line and solid-background polygons show the edge-based and node-based partitions of the network, respectively. More details and alternative partitioning schemes are described in the main text.

## 4. Discussion

### 4.1. Data Mining Genomic and Transcriptomic Resources: Sequence Quality

The results of studies exploiting genomic and transcriptomic resources depend on their quality and curatorial status, regardless of how complex and efficient the bioinformatic approaches used to extract this information are [57,58]. Usually, the scale and complexity of studies using “big data” prevent end-user control of their quality [59], and data may include unnoticed errors (e.g., incorrect taxonomic assignments or shifts in reading frames) or may have escaped objective quality filters (e.g., low sequence quality or assembly problems). Here, we used several public databases of annotated sequence data, including GenBank, FlyBase, modENCODE, OrthoDB, and BioGRID, as well as the partially released 1KITE database. Each may have contributed particular biases to the results, but the amount of data was still amenable to manual control of the different analytical steps, allowing for

the recognition of problems and for hopefully avoiding them by iterative analytical exploration and filtering of the data.

The first challenge we had to address after mining the sequence data, and before all analyses, was filtering what we interpreted as noisy sequence data or suspicious annotations in the data sets. Sequence quality was a major concern when using data directly mined from sequence repositories. Thus, we identified, through iterative assessment, two main criteria for the total or partial removal of potentially noisy data. These were (i) long autapomorphic insertions in amino acid sequences which may result from unrecognized introns and (ii) highly divergent, unalignable regions, typically at the ends of sequences, due to compensated nucleotide gains/losses in that part of the sequence, locally affecting the reading frame. Reiterated multiple sequence alignments also allowed recomposing the proteins that appeared in OrthoDB as non-overlapping fragments for some taxa into a single sequence. However, when this situation affected duplicated genes, there was a risk of joining fragments of non-orthologous proteins, which we addressed by using phylogenetic trees to inform manual curation [29]. We gained additional insight into the aforementioned problems by merging annotated and curated amino acid sequence data from OrthoDB with translated nucleotide sequence data from 1KITE beetle transcriptomes. Some of the latter sequences showed precisely the same translation problems affecting homology as were found for the insect protein data, and they were filtered according to the same criteria specified above.

#### 4.2. Data Mining Genomic and Transcriptomic Resources: Orthology Assessment

Orthology assessment was particularly crucial in the examination of beetle data mined directly from raw transcriptomes, and here, this assessment was particularly important because orthology provided our best hypothesis for conserved gene function. For most EOGs of interest for which we searched the transcriptomes, the pipeline yielded a phylogenetically cohesive group of potentially orthologous sequences with their paralogs when they were present in the transcriptome. The efficiency of Orthograph in this respect was demonstrated when mistakes were made. For example, a bad specification of the EOG corresponding to the gene *Pen* initially resulted in predicted beetle orthologs for one of the other importin-alpha genes in insects, which could be identified and corrected in our iterative phylogenetic approach. For six genes, however, the analyses picked up at least two paralogs. Two corresponded to *Act5C* and *Cyt-c-d*, which we already described as challenging to separate in the respective duplicated copies, even using phylogenies. The other four are more difficult to explain, and recognizing them required phylogenetically informed decisions; they were removed from the analyses a posteriori. Of these, two were genes for which we revealed duplications in beetles, *Dronc* and *gish* (for the latter, we found the beetle-specific paralog only in *Xylobiops* [Bostrichidae]). The other two were *Npc1a*, for which the correct sperm individualization ortholog was identified in 16 beetle transcriptomes and its paralog *Npc1b* in *Lepicerus* sp., and *klhl10*, for which the copy missing in Diptera was found in *Micromalthus debilis* and *Lamprohiza splendidula*. For all of these genes, we have strong evidence hinting at them being duplicated in the beetle genomes, yet we retrieved one of the copies in most species and the other copy in one or just a few transcriptomes. If these genes are indeed duplicated, the reason why both copies were not found consistently in all beetle transcriptomes may be related to how the program Orthograph works, i.e., retrieving a single best reciprocal hit, the putative ortholog. In these circumstances, and analogously to ranked results of BLAST searches, the correct, biologically meaningful sequence may be missed after yielding a suboptimal hit, perhaps because of sequence quality and/or length issues or the absence of the ortholog of interest in some of the transcriptomes.

#### 4.3. Evolutionary Dynamics of Sperm Individualization Genes

All qualitative traits that were used to rank sperm individualization genes in insects were statistically independent. This implies that, at least for this subset of genes, some evolutionary predictions do not apply, including the association of sex-biased gene expression with an origin

attributed to gene duplications [18]. Apart from duplications, we also recorded gene losses, because it has been hypothesized that the rate of turnover (i.e., lack of 1:1 orthology) for sex-biased—particularly male-biased—genes may be higher than for other genes [60,61]. Among sperm individualization genes, we found lineage-specific losses for both biased and unbiased genes without statistical differences between groups (chi-sq = 2.4529,  $p = 0.1450$ ), and our phylogenetic analyses, in fact, show that preservation and genomic dosage of sperm individualization genes are generally highly conserved across the Insecta despite their long evolutionary history.

In our analysis of the correlation between the different ways in which we ranked sperm individualization genes and their inferred evolutionary rates, only two instances of statistically significant differences were obtained. The first relates to the overall homogeneity of substitution rates both for insect amino acid and for beetle nucleotide sequence data (even if with opposite signs). The second and most interesting, considering the deep evolutionary time considered and the assumption of conservation of gene functionality across this time scale, was for sex-biased genes, which had different and significantly higher evolutionary rates than unbiased genes. The fact that sex-biased genes, and, more specifically, male-biased genes, evolve more rapidly than unbiased genes is a well-known general evolutionary pattern documented from a diversity of organisms [5,60–69]. However, it is surprising that this signature is still present across some 400 million years of evolution when it remains unclear whether gene functionality and sex bias in their expression have been conserved. If these features changed during the course of evolution, it is still possible that faster rates in this case could be related to other expression features, such as tissue specificity and narrow expression profiles [65]. Indeed, faster rates of evolution associated with sex-biased expression have been explained as the result of several potential causes, including participation in specific processes such as spermatogenesis [62,69], activation in reproductive tissues relative to genes expressed in several tissues [13,70], linkage to the homogametic sex chromosome [13,70], relatively low levels of expression [71], or circumscription to specific stages of development [5]. These correlations are far from universal, and there are exceptions to each of the proposed patterns [5,68,69], much depending on the organism under study but also on their life histories. For example, female mating behavior in different species of *Anopheles* [Diptera: Culicidae]—some species of which are polyandrous, while others mate once in their lifetime—may have different impacts on sperm competition and selection and, consequently, on the evolutionary dynamics of sperm-related genes [69]. Moreover, while these factors could potentially lead to faster rates of evolution in sex-biased genes, protein–protein interactions could effectively constrain them [72], a possibility that will be discussed below.

#### 4.4. Evolutionary Dynamics of Interacting Sperm Individualization Genes

Genetic interactions act as a dominant force explaining evolutionary rates, and the nature and type of interaction may prevail over other factors, such as the characteristics of gene expression [73]. There are hypotheses on how these two features may interact, such as the expected negative correlation between the number of protein interactions and evolutionary rates, or the proposition that interacting proteins should evolve at similar rates [74]. The micro- and macroevolutionary analyses of the effect of these interactions have facilitated significant advances in our understanding of these processes. On the one hand, our knowledge on the structure of genetic interaction networks, also for non-model organisms, is more detailed. On the other hand, the development of explicit, quantitative methods allows us to evaluate the architecture and properties of the networks relative to the biological features of their elements, particularly in the case of metabolic networks [75–78].

Among typical macroevolutionary patterns related to the protein–protein interaction network structure, it has been proposed that duplicated genes tend to be more highly connected in such networks [77]. The sperm individualization network shows an area that concentrates duplicated and relatively highly connected proteins (e.g., *Dredd*, *Dronc*, and *skap*); however, there was no statistical support for a correlation between these features. Correlations were found, nonetheless, for evolutionary rates when the undirected network was bipartitioned, adding statistical support to the intuitive notion of



faster-evolving genes and proteins (*blanks*, *Dark*, *Dredd*, *Dronc*, *Duba*, and *Fadd*) appearing concentrated in one region of the network. Furthermore, we found a positive correlation between rank-based centrality and nucleotide substitution rates for beetles. In general, the opposite trend tends to be the norm, and highly connected genes usually show slower rates of evolution, maybe because the protein function depends on more topological interactions with other proteins, which constrain the possibility of change [74,75]. However, this is a controversial topic, and other examples of faster-evolving core proteins in an interaction network exist, such as the analysis of transcriptional networks in yeast [79]. In any case, it is too early to draw conclusions about the evolutionary trends in sperm individualization genes. Significant results were only obtained for beetles, for which we lack empirical evidence of the same gene interactions known in *Drosophila*, and different evolutionary dynamics seem to operate depending on the overall function of the network. This highlights the necessity for further research in beetles beyond the model organism *T. castaneum*.

The lack of significant or consistent results between analyses employing insect amino acid and beetle nucleotide sequence data may be explained, in part, by the partial view of the actual interactions in which sperm individualization genes participate. It is possible that the real nature and number of these interactions is not captured by the necessarily crude correction applied here (i.e., total number of receiving edges in the interactome). The structural measures obtained from the interaction network are intrinsic and the represented network of interactions is not isolated; therefore, these measures can show some biases [75]. A poorly connected node in the sperm individualization network can have many connections to other functional domains of the cell. For example, the proteins *Fadd* and *Mer* physically interact with the products of three and four other sperm individualization genes, respectively; however, in the complete interactome of *Drosophila*, they are known to interact physically or genetically with 100 and 165 other proteins, respectively. As already mentioned, another possibility is that the interaction network described for *Drosophila* is not universal for insects, totally or partially, and that the enforced topology is unable to capture evolutionary constraints for these genes in insects, or that the actual evolutionary dynamics of beetles are different from general trends in insects. Nevertheless, we tried to find intrinsic patterns that could be associated with the coordination of the genes of interest in a specific function, and at least in the case of beetles, there could be a signature worth exploring from a functional point of view.

While we identified statistically significant differences in the rate of amino acid substitution in insects depending on hypothesized sex-biased expression, the study of nucleotide substitution rates in beetles for the same genes did not reveal any significant pattern. A somewhat reverse pattern was obtained in our exploration of evolutionary rates constrained by the architecture of a hypothesized network of interaction, wherein mainly nucleotide substitution rates of beetles showed some correlation with this architecture. This apparent contradiction and the complexity of the factors involved in explaining evolutionary rates make it difficult to fully explain these patterns satisfactorily. Before we can do that, we need more in-depth insight into the temporal and spatial expression profiles, effective function, genetic interactions, and pleiotropic effects of these genes in every single species, but also to incorporate information on their life history, which is likely to influence their evolutionary dynamics.

**Supplementary Materials:** The following are available online at <http://www.mdpi.com/2073-4425/10/10/776/s1>, File S1. Maximum likelihood trees based on the amino acid alignments of different sperm individualization proteins in insects. File S2. Bayesian inference trees based on the amino acid alignments of different sperm individualization proteins in insects. File S3. Maximum likelihood trees based on the nucleotide alignments of different sperm individualization genes in beetles. File S4. Bayesian inference trees based on the nucleotide alignments of different sperm individualization genes in beetles.

**Author Contributions:** J.G.-Z. conceived the study, analyzed and curated the data, interpreted the results, and wrote the original draft of the manuscript; D.D.K. and X.Z. contributed data as part of the 1KITE initiative; H.I.V.-R., C.M., and M.P. mined the data; J.G.-Z. and C.M. wrote the manuscript and covered publication expenses; M.P., D.D.K., and X.Z. contributed to the manuscript. All authors approved the manuscript.

**Funding:** This study was possible thanks to the project CGL2011-23820/BOS of the Spanish Ministry of Science and Innovation led by JGZ, which also included a predoctoral scholarship (BES-2012-051908) as well as two training stays (EEBB-I-14-08654 and EEBB-I-16-11559) at the Zoological Research Museum Alexander Koenig

(Bonn, Germany), funded by the Spanish Ministry of Economy and Competitiveness and enjoyed by HIVR. One of us (CM) and Bernhard Misof hosted these stays and the support of the later is much appreciated, also in his role as one of the 1KITE leaders. Indeed, this study uses data from the 1KITE consortium ([www.1kite.org](http://www.1kite.org)), which was supported by the China National Genebank and Beijing Genomics Institute (Shenzhen). We are especially grateful to the 1KITE beetle group for granting access to partially unpublished data, particularly to Kai Schütte (Hamburg, Germany), Eric Anton (Jena, Germany), Hermes Escalona and Adam Ślipiński (Canberra, Australia), Dirk Ahrens (Bonn, Germany) and Michael Balke (Munich, Germany), who provided specimens or tissue for 1KITE beetle transcriptomes, and to Alexander Donath, Lars Podsiadlowski, Shanlin Liu, Guanliang Meng and Karen Meusemann for managing and making accessible 1KITE data and accompanying information that we used for this study. MP was funded by the Leibniz Graduate School on Genomic Biodiversity Research (GBR) and by the German Research Foundation (DFG, grant MI 649/16-1).

**Conflicts of Interest:** The authors declare no conflict of interest.

## References

1. Khaitovich, P.; Hellmann, I.; Enard, W.; Nowick, K.; Leinweber, M.; Franz, H.; Weiss, G.; Lachmann, M.; Pääbo, S. Parallel patterns of evolution in the genomes and transcriptomes of humans and chimpanzees. *Science* **2005**, *309*, 1850–1854. [[CrossRef](#)] [[PubMed](#)]
2. Ranz, J.M.; Castillo-Davis, C.I.; Meiklejohn, C.D.; Hartl, D.L. Sex-dependent gene expression and evolution of the *Drosophila* transcriptome. *Science* **2003**, *300*, 1742–1745. [[CrossRef](#)] [[PubMed](#)]
3. Graveley, B.R.; Brooks, A.N.; Carlson, J.W.; Duff, M.O.; Landolin, J.M.; Yang, L.; Artieri, C.G.; van Baren, M.J.; Boley, N.; Booth, B.W.; et al. The developmental transcriptome of *Drosophila melanogaster*. *Nature* **2011**, *471*, 473–479. [[CrossRef](#)] [[PubMed](#)]
4. Singh, R.; Jagadeeshan, S. Sex and speciation: *Drosophila* reproductive tract proteins—Twenty five years later. *Int. J. Evol. Biol.* **2012**, *2012*, 191495. [[CrossRef](#)]
5. Perry, J.C.; Harrison, P.W.; Mank, J.E. The ontogeny and evolution of sex-biased gene expression in *Drosophila melanogaster*. *Mol. Biol. Evol.* **2014**, *31*, 1206–1219. [[CrossRef](#)]
6. Baker, D.A.; Nolan, T.; Fischer, B.; Pinder, A.; Crisanti, A.; Russell, S. A comprehensive gene expression atlas of sex- and tissue-specificity in the malaria vector, *Anopheles gambiae*. *BMC Genom.* **2011**, *12*, 296. [[CrossRef](#)]
7. Prince, E.G.; Kirkland, D.; Demuth, J.P. Hyperexpression of the X chromosome in both sexes results in extensive female bias of X-linked genes in the flour beetle. *Genome Biol. Evol.* **2010**, *2*, 336–346. [[CrossRef](#)]
8. Parsch, J.; Ellegren, H. The evolutionary causes and consequences of sex-biased gene expression. *Nat. Rev. Genet.* **2013**, *14*, 83–87. [[CrossRef](#)]
9. Telonis-Scott, M.; Kopp, A.; Wayne, M.L.; Nuzhdin, S.V.; McIntyre, L.M. Sex-specific splicing in *Drosophila*: Widespread occurrence, tissue specificity and evolutionary conservation. *Genetics* **2009**, *181*, 421–434. [[CrossRef](#)]
10. Hartmann, B.; Castelo, R.; Miñana, B.; Peden, E.; Blanchette, M.; Rio, D.C.; Singh, R.; Valcárcel, J. Distinct regulatory programs establish widespread sex-specific alternative splicing in *Drosophila melanogaster*. *RNA* **2011**, *17*, 453–468. [[CrossRef](#)]
11. Meisel, R.P.; Malone, J.H.; Clark, A.G. Disentangling the relationship between sex-biased gene expression and X-linkage. *Genome Res.* **2012**, *22*, 1255–1265. [[CrossRef](#)] [[PubMed](#)]
12. Lee, H.; Cho, D.Y.; Whitworth, C.; Eisman, R.; Phelps, M.; Roote, J.; Kaufman, T.; Cook, K.; Russell, S.; Przytycka, T.; et al. Effects of gene dose, chromatin, and network topology on expression in *Drosophila melanogaster*. *PLoS Genet.* **2016**, *12*, e1006295. [[CrossRef](#)] [[PubMed](#)]
13. Ranz, J.M.; Parsch, J. Newly evolved genes: Moving from comparative genomics to functional studies in model systems. *Bioessays* **2012**, *34*, 477–483. [[CrossRef](#)] [[PubMed](#)]
14. Gallach, M.; Domingues, S.; Betrán, E. Gene duplication and the genome distribution of sex-biased genes. *Intl. J. Evol. Biol.* **2011**, *2011*, 989438. [[CrossRef](#)] [[PubMed](#)]
15. Chen, S.; Krinsky, B.H.; Long, M.y. New genes as drivers of phenotypic evolution. *Nat. Rev. Genet.* **2013**, *14*, 645–660. [[CrossRef](#)] [[PubMed](#)]
16. Zhang, Z.; Hambuch, T.M.; Parsch, J. Molecular evolution of sex-biased genes in *Drosophila*. *Mol. Biol. Evol.* **2004**, *21*, 2130–2139. [[CrossRef](#)] [[PubMed](#)]
17. Pröschel, M.; Zhang, Z.; Parsch, J. Widespread adaptive evolution of *Drosophila* genes with sex-biased expression. *Genetics* **2006**, *174*, 893–900. [[CrossRef](#)]

18. Ellegren, H.; Parsch, J. The evolution of sex-biased genes and sex-biased gene expression. *Nat. Rev. Genet.* **2007**, *8*, 689–698. [[CrossRef](#)]
19. Haerty, W.; Jagadeeshan, S.; Kulathinal, R.J.; Wong, A.; Ram, K.R.; Sirot, L.K.; Levesque, L.; Artieri, C.G.; Wolfner, M.F.; Civetta, A.; et al. Evolution in the fast lane: Rapidly evolving sex-related genes in *Drosophila*. *Genetics* **2007**, *177*, 1321–1335. [[CrossRef](#)]
20. Yang, Y.; Smith, S.A. Orthology inference in nonmodel organisms using transcriptomes and low-coverage genomes: Improving accuracy and matrix occupancy for phylogenomics. *Mol. Biol. Evol.* **2014**, *31*, 3081–3092. [[CrossRef](#)]
21. Sjölander, K. Phylogenomic inference of protein molecular function: Advances and challenges. *Bioinformatics* **2004**, *20*, 170–179. [[CrossRef](#)] [[PubMed](#)]
22. Vizán-Rico, H.I.; Gómez-Zurita, J. Testis-specific RNA-Seq of *Calligrapha* (Chrysomelidae) as a transcriptomic resource for male-biased gene inquiry in Coleoptera. *Mol. Ecol. Res.* **2017**, *17*, 533–545. [[CrossRef](#)] [[PubMed](#)]
23. Grath, S.; Parsch, J. Sex-biased gene expression. *Ann. Rev. Genet.* **2016**, *50*, 29–44. [[CrossRef](#)] [[PubMed](#)]
24. Parisi, M.; Nuttall, R.; Edwards, P.; Minor, J.; Naiman, D.; Lü, J.; Doctolero, M.; Vainer, M.; Chan, C.; Malley, J.; et al. A survey of ovary-, testis-, and soma-biased gene expression in *Drosophila melanogaster* adults. *Genome Biol.* **2004**, *5*, R40. [[CrossRef](#)] [[PubMed](#)]
25. Fabrizio, J.J.; Hime, G.; Lemmon, S.K.; Bazinet, C. Genetic dissection of sperm individualization in *Drosophila melanogaster*. *Development* **1998**, *125*, 1833–1843.
26. Fuller, M.T. Spermatogenesis. In *The Development of Drosophila melanogaster*; Bate, M., Arias, A.M., Eds.; Cold Spring Harbor Laboratory Press: New York, NY, USA, 1993; pp. 71–147.
27. Celniker, S.E.; Dillon, L.A.; Gerstein, M.B.; Gunsalus, K.C.; Henikoff, S.; Karpen, G.H.; Kellis, M.; Lai, E.C.; Lieb, J.D.; MacAlpine, D.M.; et al. Unlocking the secrets of the genome. *Nature* **2009**, *459*, 927–930. [[CrossRef](#)]
28. Kriventseva, E.V.; Tegenfeldt, F.; Petty, T.J.; Waterhouse, R.M.; Simão, F.A.; Pozdnyakov, I.A.; Ioannidis, P.; Zdobnov, E.M. OrthoDB v8: Update of the hierarchical catalog of orthologs and the underlying free software. *Nucleic Acids Res.* **2015**, *43*, D250–D256. [[CrossRef](#)]
29. Gabaldón, T. Large-scale assignment of orthology: Back to phylogenetics? *Genome Biol.* **2008**, *9*, 235. [[CrossRef](#)]
30. Carbon, S.; Ireland, A.; Mungall, C.J.; Shu, S.Q.; Marshall, B.; Lewis, S. AmiGO: Online access to ontology and annotation data. *Bioinformatics* **2009**, *25*, 288–289. [[CrossRef](#)]
31. Gramates, L.S.; Marygold, S.J.; dos Santos, G.; Urbano, J.-M.; Antonazzo, G.; Matthews, B.B.; Rey, A.J.; Tabone, C.J.; Crosby, M.A.; Emmert, D.B.; et al. FlyBase at 25: Looking to the future. *Nucleic Acids Res.* **2017**, *45*, D663–D671. [[CrossRef](#)]
32. Kalderimis, A.; Lyne, R.; Butano, D.; Contrino, S.; Lyne, M.; Heimbach, J.; Hu, F.; Smith, R.; Stěpán, R.; Sullivan, J.; et al. InterMine: Extensive web services for modern biology. *Nucleic Acids Res.* **2014**, *42*, W468–W472. [[CrossRef](#)] [[PubMed](#)]
33. Zdobnov, E.M.; Tegenfeldt, F.; Kuznetsov, D.; Waterhouse, R.M.; Simão, F.A.; Panagiotis, I.; Seppey, M.; Loetscher, A.; Kriventseva, E.V. OrthoDB v9.1: Cataloguing evolutionary and functional annotations for animal, fungal, plant, archaeal, bacterial and viral orthologs. *Nucleic Acids Res.* **2017**, *45*, D744–D749. [[CrossRef](#)] [[PubMed](#)]
34. Petersen, M.; Meusemann, K.; Donath, A.; Dowling, D.; Liu, S.; Peters, R.S.; Podsiadlowski, L.; Vasilikopoulos, A.; Zhou, X.; Misof, B.; et al. Orthograph: A versatile tool for mapping coding nucleotide sequences to clusters of orthologous genes. *BMC Bioinform.* **2017**, *18*, 111. [[CrossRef](#)] [[PubMed](#)]
35. Attrill, H.; Falls, K.; Goodman, J.L.; Millburn, G.H.; Antonazzo, G.; Rey, A.J.; Marygold, S.J. FlyBase Consortium. FlyBase: Establishing a Gene Group resource for *Drosophila melanogaster*. *Nucleic Acids Res.* **2016**, *44*, D786–D792. [[CrossRef](#)] [[PubMed](#)]
36. Kim, H.S.; Murphy, T.; Xia, J.; Caragea, D.; Park, Y.; Beeman, R.W.; Lorenzen, M.D.; Butcher, S.; Manak, J.R.; Brown, S.J. BeetleBase in 2010: Revisions to provide comprehensive genomic information for *Tribolium castaneum*. *Nucleic Acids Res.* **2010**, *38*, D437–D442. [[CrossRef](#)] [[PubMed](#)]
37. Nygaard, S.; Zhang, G.; Schiott, M.; Li, C.; Wurm, Y.; Hu, H.F.; Zhou, J.J.; Ji, L.; Qiu, F.; Rasmussen, M.; et al. The genome of the leaf-cutting ant *Acromyrmex echinatior* suggests key adaptations to advanced social life and fungus farming. *Genome Res.* **2011**, *21*, 1339–1348. [[CrossRef](#)] [[PubMed](#)]

38. Elsik, C.G.; Tayal, A.; Diesh, C.M.; Unni, D.R.; Emery, M.L.; Nguyen, H.N.; Hagen, D.E. Hymenoptera Genome Database: Integrating genome annotations in HymenopteraMine. *Nucleic Acids Res.* **2016**, *44*, D793–D800. [[CrossRef](#)]
39. Katoh, S. MAFFT multiple sequence alignment software version 7: Improvements in performance and usability. *Mol. Biol. Evol.* **2013**, *30*, 772–780. [[CrossRef](#)]
40. Lefort, V.; Longueville, J.E.; Gascuel, O. SMS: Smart Model Selection in PhyML. *Mol. Biol. Evol.* **2017**, *34*, 2422–2424. [[CrossRef](#)]
41. Guindon, S.; Dufayard, J.-F.; Lefort, V.; Anisimova, M.; Hordijk, W.; Gascuel, O. New algorithms and methods to estimate maximum-likelihood phylogenies: Assessing the performance of PhyML 3.0. *Syst. Biol.* **2010**, *59*, 307–321. [[CrossRef](#)]
42. Drummond, A.J.; Suchard, M.A.; Xie, D.; Rambaut, A. Bayesian phylogenetics with BEAUti and the BEAST 1.7. *Mol. Biol. Evol.* **2012**, *29*, 1969–1973. [[CrossRef](#)] [[PubMed](#)]
43. Ronquist, F.; Klopfstein, S.; Vilhelmsen, L.; Schulmeister, S.; Murray, D.L.; Rasnitsyn, A.P. A total-evidence approach to dating with fossils, applied to the early radiation of the Hymenoptera. *Syst. Biol.* **2012**, *61*, 973–999. [[CrossRef](#)] [[PubMed](#)]
44. Bertone, M.A.; Courtney, G.W.; Wiegmann, B.M. Phylogenetics and temporal diversification of the earliest true flies (Insecta: Diptera) based on multiple nuclear genes. *Syst. Ent.* **2008**, *33*, 668–687. [[CrossRef](#)]
45. Drummond, A.J.; Ho, S.Y.; Phillips, M.J.; Rambaut, A. Relaxed phylogenetics and dating with confidence. *PLoS Biol.* **2006**, *4*, e88. [[CrossRef](#)]
46. Rambaut, A.; Suchard, M.A.; Xie, D.; Drummond, A.J. Tracer v1.6. 2014. Available online: <http://beast.bio.ed.ac.uk/Tracer> (accessed on 10 March 2015).
47. Mann, H.B.; Whitney, D.R. On a test of whether one of two random variables is stochastically larger than the other. *Ann. Math. Stat.* **1947**, *18*, 50–60. [[CrossRef](#)]
48. R Core Team. *R: A Language and Environment for Statistical Computing*; R Foundation for Statistical Computing: Vienna, Austria, 2016.
49. Weiss, N.A. wPerm. Permutation Tests. R package version 1.0.1. 2015. Available online: <https://CRAN.R-project.org/package=wPerm> (accessed on 15 February 2018).
50. Stark, C.; Breitkreutz, B.-J.; Reguly, T.; Boucher, L.; Breitkreutz, A.; Tyers, M. BioGRID: A general repository for interaction datasets. *Nucleic Acids Res.* **2006**, *34*, D535–D539. [[CrossRef](#)]
51. Csardi, G.; Nepusz, T. The igraph software package for complex network research. *Int. J. Complex. Syst.* **2006**, *1695*, 1–9.
52. Blondel, V.; Guillaume, J.-L.; Lambiotte, R.; Lefebvre, E. Fast unfolding of communities in large networks. *J. Stat. Mech.* **2008**, *10*, P10008. [[CrossRef](#)]
53. Brandes, U.; Delling, D.; Gaertler, M.; Görke, R.; Hofer, M.; Nikoloski, Z.; Wagner, D. On modularity clustering. *IEEE Trans. Knowl. Data Eng.* **2008**, *20*, 172–188. [[CrossRef](#)]
54. Brandes, U. A faster algorithm for betweenness centrality. *J. Math. Sociol.* **2001**, *25*, 163–177. [[CrossRef](#)]
55. Kruskal, W.H.; Wallis, W.A. Use of ranks in one-criterion variance analysis. *J. Am. Stat. Assoc.* **1952**, *47*, 583–621. [[CrossRef](#)]
56. Misof, B.; Liu, S.I.; Meusemann, K.; Peters, R.S.; Donath, A.; Mayer, C.; Frandsen, P.B.; Ware, J.; Flouri, T.; Beutel, R.G.; et al. Phylogenomics resolves the timing and pattern of insect evolution. *Science* **2014**, *346*, 763–767. [[CrossRef](#)] [[PubMed](#)]
57. Pruitt, K.D.; Tatusova, T.; Maglott, D.R. NCBI Reference Sequence (RefSeq): A curated non-redundant sequence database of genomes, transcripts and proteins. *Nucleic Acids Res.* **2005**, *33*, D501–D504. [[CrossRef](#)] [[PubMed](#)]
58. Klimke, W.; O'Donovan, C.; White, O.; Brister, J.R.; Clark, K.; Fedorov, B.; Mizrahi, I.; Pruitt, K.D.; Tatusova, T. Solving the problem: Genome annotation standards before the data deluge. *Stand. Genom. Sci.* **2011**, *5*, 168–193. [[CrossRef](#)]
59. Holzinger, A.; Dehmer, M.; Jurisica, I. Knowledge Discovery and interactive data mining in Bioinformatics -state-of-the-art, future challenges and research directions. *BMC Bioinform.* **2014**, *15*, I1. [[CrossRef](#)]
60. Zhang, Y.; Sturgill, D.; Parisi, M.; Kumar, S.; Oliver, B. Constraint and turnover in sex-biased gene expression in the genus *Drosophila*. *Nature* **2007**, *450*, 233–238. [[CrossRef](#)]
61. Assis, R.; Zhou, Q.; Bachtrog, D. Sex-biased transcriptome evolution in *Drosophila*. *Genome Biol. Evol.* **2012**, *4*, 1189–1200. [[CrossRef](#)]



62. Torgerson, D.G.; Kulathinal, R.J.; Singh, R.S. Mammalian sperm proteins are rapidly evolving: Evidence of positive selection in functionally diverse genes. *Mol. Biol. Evol.* **2002**, *19*, 1973–1980. [[CrossRef](#)]
63. Jagadeeshan, S.; Singh, R.S. Rapidly evolving genes of *Drosophila*: Differing levels of selective pressure in testis, ovary, and head tissues between sibling. *Mol. Biol. Evol.* **2005**, *22*, 1793–1801. [[CrossRef](#)]
64. Zhang, Z.; Parsch, J. Positive correlation between evolutionary rate and recombination rate in *Drosophila* genes with male-biased expression. *Mol. Biol. Evol.* **2005**, *22*, 1945–1947. [[CrossRef](#)]
65. Meisel, R.P. Towards a more nuanced understanding of the relationship between sex-biased gene expression and rates of protein-coding sequence evolution. *Mol. Biol. Evol.* **2011**, *28*, 1893–1900. [[CrossRef](#)] [[PubMed](#)]
66. Müller, L.; Grath, S.; von Heckel, K.; Parsch, J. Inter- and intraspecific variation in *Drosophila* genes with sex-biased expression. *Int. J. Evol. Biol.* **2012**, 963–976. [[CrossRef](#)]
67. Wang, X.; Werren, J.H.; Clark, A.G. Genetic and epigenetic architecture of sex-biased expression in the jewel wasps *Nasonia vitripennis* and *giraulti*. *Proc. Natl. Acad. Sci. USA* **2015**, *112*, E3545–E3554. [[CrossRef](#)] [[PubMed](#)]
68. Darolti, I.; Wright, A.E.; Pucholt, P.; Berlin, S.; Mank, J.E. Slow evolution of sex-biased genes in the reproductive tissue of the dioecious plant *Salix viminalis*. *Mol. Ecol.* **2018**, *27*, 694–708. [[CrossRef](#)]
69. Papa, F.; Windbichler, N.; Waterhouse, R.M.; Cagnetti, A.; D'Amato, R.; Persampieri, T.; Lawniczak, M.K.N.; Nolan, T.; Papatianos, P.A. Rapid evolution of female-biased genes among four species of *Anopheles* malaria mosquitoes. *Genome Res.* **2018**, *27*, 1536–1548. [[CrossRef](#)]
70. Grath, S.; Parsch, J. Rate of amino acid substitution is influenced by the degree and conservation of male-biased transcription over 50 myr of *Drosophila* evolution. *Genome Biol. Evol.* **2012**, *4*, 346–359. [[CrossRef](#)] [[PubMed](#)]
71. Drummond, D.A.; Bloom, J.D.; Adami, C.; Wilke, C.O.; Arnold, F.H. Why highly expressed proteins evolve slowly. *Proc. Natl. Acad. Sci. USA* **2005**, *102*, 14338–14343. [[CrossRef](#)]
72. Wong, A.; Wolfner, M.F. Evolution of *Drosophila* seminal proteins and their networks. In *Rapidly Evolving Genes & Genetic Systems*; Singh, R.S., Xu, J.P., Kulathinal, R.J., Eds.; Oxford University Press: Oxford, UK, 2012; pp. 144–152.
73. Fraser, H.B.; Wall, D.P.; Hirsh, A.E. A simple dependence between protein evolution rate and the number of protein-protein interactions. *BMC Evol. Biol.* **2003**, *3*, 11. [[CrossRef](#)] [[PubMed](#)]
74. Fraser, H.B.; Hirsh, A.E.; Steinmetz, L.M.; Scharfe, C.; Feldman, M.W. Evolutionary rate in the protein interaction network. *Science* **2002**, *296*, 750–752. [[CrossRef](#)]
75. Cork, J.M.; Purugganan, M.D. The evolution of molecular genetic pathways and networks. *Bioessays* **2004**, *26*, 479–484. [[CrossRef](#)]
76. Wagner, A. Metabolic networks and their evolution. In *Evolutionary Systems Biology*; Soyer, O.S., Ed.; Springer: New York, NY, USA, 2012; pp. 29–52.
77. Alvarez-Ponce, D.; Fares, M.A. Evolutionary rate and duplicability in the *Arabidopsis thaliana* protein-protein interaction network. *Genome Biol. Evol.* **2012**, *4*, 1263–1274. [[CrossRef](#)] [[PubMed](#)]
78. Colombo, M.; Laayouni, H.; Invergo, B.M.; Bertranpetit, J.; Montanucci, L. Metabolic flux is a determinant of the evolutionary rates of enzyme-encoding genes. *Evolution* **2013**, *68*, 605–613. [[CrossRef](#)] [[PubMed](#)]
79. Jovelin, R.; Phillips, P.C. Evolutionary rates and centrality in the yeast gene regulatory network. *Genome Biol.* **2009**, *10*, R35. [[CrossRef](#)] [[PubMed](#)]



## 11. Discusión global

El desarrollo de la presente tesis doctoral ha contribuido al conocimiento con una lista de resultados, así como con interesantes discusiones y reflexiones tanto para el presente como para investigaciones futuras. A continuación, exponemos su contenido estructurado en siete secciones, de acuerdo con los capítulos de las publicaciones anteriormente mencionadas.

### **11.1. Obtención del material fuente: secuenciación de cinco transcriptomas de cuatro especies no modelo del género *Calligrapha***

Las especies modelo juegan un papel decisivo en el avance científico y es lícito fomentar un estudio completo y multidisciplinar sobre ellas; sin embargo, no deben ni infuscar ni restar importancia al estudio de las especies restantes, y sirvan para ello los siguientes argumentos: (i) obviar el estudio de especies no modelo puede ignorar procesos interesantes, imposibles de conocer *a priori*; (ii) el estudio de funciones o elementos en especies modelo puede resultar igualmente útil en las especies nuevas e, incluso, podría revelarse que estas últimas habrían podido ser mejores modelos que aquellas; (iii) la amplitud de un abanico de especies enriquece la visión de cualquier estudio y favorece la comprensión de los elementos del mismo, siendo la comparación interespecífica necesaria para el discernimiento de patrones homólogos entre taxones, por añadidura.

Dicho lo anterior, el material fuente del que se nutre esta tesis (la secuenciación de mRNA expresado en ejemplares de cinco especies del género *Calligrapha*), ha contribuido al aumento del acervo genético público disponible, permitiendo enriquecer cualquier estudio ulterior, especialmente sobre diversidad genética.

De las más de ochenta especies que posee el género *Calligrapha* Chevrolat 1836 (Chrysomelidae) distribuidas por el íntegro continente americano, desde Norteamérica (Alaska y todas y cada una de las provincias de Canadá) hasta Argentina (Gómez-Zurita, 2005), nuestra secuenciación de cuatro especies de América del norte puede parecer escasa o poco representativa. Sin embargo, tratándose del primer estudio de secuenciación de especies del género *Calligrapha* y, dada la cantidad de secuencias que produce la NGS, el análisis de cinco transcriptomas puede considerarse una cantidad aceptable.

Podría pensarse que la elección de especies geográficamente en sus máximos distantes (Canadá-Argentina) hubiera sido excelente; sin embargo, esto no es completamente necesario, ya que: (i) la distancia geográfica no implica distancia genética *per se*; (ii) la distancia genética entre especies es suficiente para estudios comparativos de secuencia a nivel de orden o clase de especies. A modo de ejemplo, *C. philadelphica* y *C. confluens* distan ~0,720 MA (Kumar et al. 2017)).

Para la obtención óptima de RNA, a diferencia del DNA, es necesario hacer la vivisección de forma rápida y en un medio líquido que asegure la preservación del RNA (típicamente el conocido RNAlater de Qiagen). Esto convierte el

momento de recolección en un momento delicado y decisivo para la calidad de la muestra y sus posibilidades de utilizarla para fines posteriores. Además, es importante una adecuada elección del momento vital de los especímenes que se quiere analizar, ya que el mRNA se sintetiza específicamente para la traducción de aquellos genes que se requieren, degradándose a continuación. En nuestro caso, con el objetivo de describir genes involucrados en funciones reproductivas, escogimos machos adultos, en los que los tejidos y procesos reproductivos están presentes y activos.

Las contaminaciones, por lo general, son difíciles de evitar en su totalidad, debido a las condiciones prácticas en las que se realiza el muestreo. En un caso de recolección estándar, suelen estar implicados varios investigadores aislados o semiaislados en un área del campo al aire libre, con sus materiales y herramientas de muestreo esterilizadas, pero temporalmente expuestas a la intemperie. Las contaminaciones, sin embargo —tal vez porque se han convertido en omnipresentes—, se corrigen durante el proceso de ensamblado, una cuestión que se trata más adelante.

Como se documenta en los métodos del artículo primero, el total del mRNA extraído de cada muestra de testículo se secuenció en Eurofins MWG Operon (Ebersberg, Germany). La estrategia empleada (que combinaba la secuenciación de muestras de testículo vía Illumina, con la de un abdomen en 454, que se usaba como referencia de mapeo) fue sin duda una acertada solución para suplir la ausencia de genoma de referencia, el cual facilita el subsecuente ensamblado de los *reads* y es altamente recomendable. Aun así, el tiempo empleado para su ensamblaje se prolongó demasiado (varios meses) en nuestra opinión. Las ventajas de la externalización del ensamblaje conllevan al irremediable inconveniente de no poder discernir entre los retrasos debidos a la naturaleza de los datos y las estrictas necesidades económicas de una empresa (como sería, por ejemplo, el hecho de tener que esperar a llenar los ocho canales de la secuenciación).

La diferencia entre las dos estrategias de ensamblado *de novo* de los *reads* (i.e., *standard contig assembly* y *isotig-strategy*) consiste en que esta última permite el ensamblaje de los contigs isofórmicos asumiendo que constituyen —idealmente— un gen. En nuestros análisis utilizamos preferentemente las secuencias de isotigs, ya que dicha estrategia es conceptualmente convincente, elimina la redundancia de los contigs y es la que hemos utilizado con preferencia para los análisis.

## 11.2. Anotación funcional de novo de cinco transcriptomas del género *Calligrapha*

Recurrir a programas de anotación masiva como Blast2GO (Conesa et al. 2005; Conesa & Götz, 2008; Götz et al. 2008; Götz et al. 2011) es casi obligado cuando se trabaja con especies que se secuencian por primera vez, como ocurre con esta tesis, que recoge la anotación funcional *de novo* de cinco transcriptomas de las cuatro especies genéticamente desconocidas.

A modo de puntualización técnica, anotamos la lentitud de anotación de Blast2GO cuando es utilizado con su interfaz gráfica. Se puede (y recomendamos) corregir este inconveniente temporal sustituyendo el paso de blast en interfaz gráfica por un blast local, e importando después el archivo en .xml al programa; el resto de pasos para la anotación son bastante plausibles en cuanto a tiempo y eficacia de esta.

El condicionante hecho de no tener en posesión ningún genoma de referencia filogenéticamente cercano a nuestro género de estudio *Calligrapha*, nos obligó a recurrir al genoma de *Tribolium castaneum* (la única especie modelo existente en Coleoptera en el momento de la secuenciación) para ejercer de referencia durante la anotación. Sin embargo, esta especie, que pertenece a la familia Tenebrionidae, dista bastante de Chrysomelidae. A modo de ejemplo, véase que el tiempo promedio de divergencia entre las especies de *T. castaneum* y *Calligrapha philadelphica* es de 233 Ma, y su tiempo estimado de divergencia, de 208 Ma (162 - 254 Ma) (Kumar et al. 2017). Las diferencias acumuladas entre dos genomas con ~200 Ma de historia evolutiva independiente son, propiamente, dignas de atención. Además, se sumó a esta argumentación la baja calidad del genoma de *T. castaneum* y el número, cantidad y calidad de herramientas a su disposición, que no son comparables en ningún momento a Flybase (Attril et al. 2016).

Intentar la anotación del 100% de los transcritos secuenciados hubiera sido excelente, pero habría sido, sencillamente, imposible; y, además, habría carecido de sentido. Establecer un umbral de longitud de secuencia en anotación de secuencias procedentes de NGS es absolutamente necesario, puesto que el número de secuencias extremadamente cortas alcanza cifras altísimas. Por otro lado, la pérdida de información puede considerarse despreciable, ya que estas secuencias contienen poca información génica (a excepción de los microRNAs, cuyo valor es indudable): en ocasiones corresponden a fragmentos sin éxito de mapeo, o a secuencias redundantes, y un número considerable de ellas están compuestas de una sola base.

Nuestro éxito del 32,8-44,6% en la anotación funcional (en particular: 11.409, 10.490, 10.423, 11.656 y 11.038 secuencias respectivamente para cada transcriptoma de *C. confluens*, *C. floridana*, *C. multipunctata*, *C. philadelphia* (PA) y *C. philadelphia* (QC)) se encuentra dentro de la normalidad si se lo compara con estudios similares (REFS). Sin embargo, es interesante comentar algunas variables que influyen y pueden matizar ciertas diferencias.

Aunque realizamos algunas pruebas con blastn, recomendamos el algoritmo blastx para búsquedas de similitud interespecífica, ya que la conservación de la secuencia de aminoácidos es mucho mayor que la nucleotídica, sin perjuicio de la funcionalidad.

10E-5 es el E-value más utilizado en búsquedas de blast, y tal vez por ello hay una tendencia general a ajustarse a él para cualquier análisis. El trivial compromiso entre E-value y especificidad permite cierto margen, empero: el 10E-3 que utilizamos nosotros en los blasts de anotación funcional seguramente aumentó el número de secuencias recuperadas por similitud de secuencia, a la vez que disminuyó su especificidad, y consiguiendo probablemente un resultado menos robusto pero dueño de más representatividad.

Fue sorprendente que los dos primeros tercios aproximadamente de blast funcional resultaran en secuencias de *T. castaneum* (que aparece en primera posición), seguido de *Dendroctonus ponderosae* (que aparece en segunda), en lugar de la especie más cercana filogenéticamente: el escarabajo de la patata *Leptinotarsa decemlineata*, el cual no aparece en el listado de *hits* hasta llegado el puesto número 23. Este hecho no puede explicarse sino por el fuerte sesgo que existe inherentemente en la base de datos de referencia: la *database nt* en ese momento concreto, previo a la secuenciación del escarabajo de la patata.

En puestos intermedios entre los hits de *D. ponderosae* y *L. decemlineata* aparecen especies de insectos, como el pulgón *Acyrtosiphon pisum*, la avispa parasitaria *Nasonia vitripennis*, o el gusano de seda *Bombix mori*, pero también, nematodos como *Trichinella spiralis* o incluso vertebrados como el pez *Danio rerio*.

Errores de anotación aparte, es evidente el gigantesco sesgo taxonómico de las bases de datos actuales, en que las especies modelo ocupan un espacio muy importante. Como efecto colateral, estas acumulan recursos que desfavorecen el estudio de las restantes. Solo cuando los bancos de secuencias sean representativos de la diversidad del planeta podremos hacer estudios con mayor facilidad y certeza.

### 11.2.1. Comparación intra- e intertranscriptómica de la redundancia funcional

Nuestro diseño experimental está basado en modo de réplicas experimentales — cuatro réplicas interespecíficas: *C. confluens*, *C. floridana*, *C. multipunctata* y *C. philadelphia*, y dos intraespecíficas: *C. philadelphia* (PA) y *C. philadelphia* (QC)—, por lo que la naturaleza de nuestros datos admite las comparaciones inter e intraespecíficas de las muestras. Si bien, estas comparaciones inter- e intraespecíficas de los transcriptomas se pueden abordar desde diferentes perspectivas.

Es importante comentar que a través de la secuenciación masiva NGS se obtiene una gran cantidad de secuencias redundantes, pero que, en nuestro trabajo, al utilizar los isotigs, redujimos el número de contigs redundantes durante el proceso de ensamblaje.

Para inferir análisis de redundancia funcional, la clasificación organizada del Gene Ontology Consortium (2000) es idóneo; dado que cada transcrito (isotig), una vez procesado por el *pipeline* de Blast2GO (Conesa et al. 2005), resulta asociado a uno o más términos de *gene ontology* (GOs), podemos realizar las comparaciones de estos términos.

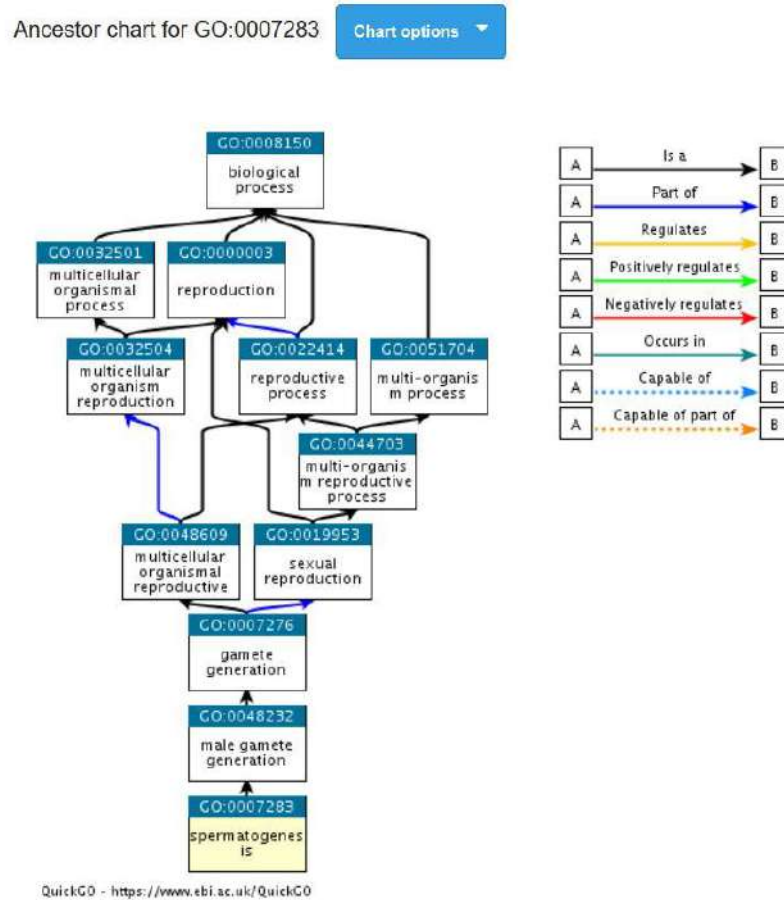
Para facilitar el manejo de datos de Gene Ontology, el navegador QuickGO-EMBL-EBI (Huntley et al. 2014; accesible en <https://www.ebi.ac.uk/QuickGO/>) resulta de gran utilidad al proveer rápidamente un gráfico de los ancestros y los *child terms*, y las relaciones entre ellos con flechas de diferente naturaleza: *is a*; *is part of*; *regulates*; *positively regulates*; *negatively regulates*; *occurs in*; *capable of*; *capable of part of*.

Pero basar las comparaciones en los términos de ontología génica tiene evidentes limitaciones. Para empezar, los términos de GO son jerárquicos y anidados, de modo que algunos de ellos engloban a muchos otros.

Se suma a esto que los términos de GO de procesos biológicos (BP) pueden resultar semánticamente ambiguos. Así, en un término *reproduction* (GO:0000003), tendrían cabida muchas funciones que no fueran meramente reproductivas, como los receptores olfativos o los receptores visuales para el reconocimiento, con fines reproductivos, de individuos de la propia especie; en definitiva, cualquier gen que interviniera en la eficacia reproductiva de un individuo sería digna de inclusión.



La figura muestra el término *spermatogenesis* (GO:0007283) de Gene Ontology y sus relaciones jerárquicas y anidadas con sus términos relativos.



**Fig. 4.** Relaciones entre términos de Gene Ontology (GOs) en torno al término GO *spermatogenesis* (GO:0007283)

Otra de las opciones para el análisis comparativo intra e interespecífico es utilizar métodos de similitud de secuencia. El archiutilizado blast continúa siendo la herramienta más útil para este fin. El uso de blast en modo local agiliza cualquier análisis de forma rápida y eficaz (ver apartado 4.3.2., donde se desarrolla). Nuestras comparaciones intra- e interespecíficas a nivel de secuencia resultaron congruentes con lo esperado, es decir, mayor similitud intraespecíficamente y menor interespecífica, y paralela o equivalente entre estas.

### 11.3. Gene-finding en transcriptomas de especies de Insecta

#### 11.3.1. Elección de genes diana

##### 11.3.1.1. Primera aproximación

En nuestra búsqueda de referentes genómicos para identificar genes putativamente homólogos a genes diana, recurrimos a la información previa disponible para la especie de referencia en insectos, i.e., *D. melanogaster*, y para aquella que, ocupando un lugar ligeramente inferior, pero por razón de compromiso, es más cercana filogenéticamente: *T. castaneum*.

FlyBase (Attril et al. 2016) es una excelente plataforma para disponer de información de expresión génica sobre especies del género *Drosophila*, pero desafortunadamente para *T. castaneum* las prestaciones se reducen considerablemente. El análisis de la expresión génica se ve constreñida a utilizar individuales estudios, como el que utilizamos aquí: el de Prince et al. (2010), pionero en la descripción de genes sex-biased en el escarabajo de la harina.

Aunque realmente encontramos interesantes genes que desempeñan funciones típicamente masculinas (como *pkd2*, *boule*, *sry*, *sarah* o *fruitless*) en nuestros transcriptomas, no pudimos incluirlos en nuestro análisis porque no poseen una expresión sesgada exclusiva de machos o predominante en ellos.

En cambio, los genes *CG9313*, *CG10859*, *Tektin-A*, *Klp59C*, *Klp59D*, *gskt*, *tomboy40* o *TLL3B* sí que presentaban un patrón de expresión exclusivo o hiperexpresión en machos y, por tanto, los etiquetamos como candidatos para futuros estudios sobre su evolución.

##### 4.3.1.2. Segunda aproximación: representantes del Gene Ontology GO:0007291 (sperm individualization)

Decidimos confiar nuestro estudio a escala mayor a la curada base de datos Gene Ontology, la cual tiene clasificados los genes por etiquetas de función, deducidas de estudios de expresión génica. Los genes incluidos bajo el término categorizado en Gene Ontology como GO:0007291 son 72. Sin embargo, es importante tener en cuenta que una pequeña cantidad de ellos es redundante. Estos se descartaron y, en consecuencia, el número total de genes seleccionados para buscar en los transcriptomas de *Calligrapha* fue finalmente de 44.

### 11.3.2. Gene-finding en los cinco transcriptomas de Calligrapha

La herramienta de blast, cuando es utilizada de forma local es rápida y eficaz. En dos simples órdenes se obtiene el resultado de un blast deseado: (i) crear una base de datos (que hará las veces de referencia) y (ii) correr el blast (con las especificaciones que se deseen).

Ejemplo general:

```
makeblastdb -in file.fas -dbtype nucl  
blastn -db database.fas -query filequery.fas -out fileout
```

La forma más eficaz consiste en escribir en un archivo de texto plano todas las órdenes y ejecutar dicho archivo, lo que genera, en cuestión de segundos, el resultado de múltiples blasts ordenados.

Los formatos más útiles de salida fueron el outfmt 5 (que genera un archivo .xml, necesario, por ejemplo, para correr un Blast2GO local), el outfmt 6 (que es un formato tabular que muestra los parámetros resultado de blast en 13 columnas: Query id, Subject id, % identity, alignment length, mismatches, gap, openings, query start, query end, subject start, subject end, e-value and bit score), el outfmt 7 (que es un formato tabular con líneas comentario) y el outfmt 10 (que ofrece los valores separados por comas). Además, los formatos 6, 7 y 10 se pueden personalizar en cuanto a qué parámetros mostrar.

Ejemplo de búsqueda de blast con outfmt 6:

```
blastx -db all_gproducts.txt -query c500iso -evalue 1e-5 -outfmt  
6 -out tab_blastxresult_allgprod0007291_c500iso
```

Si se quiere cuantificar de forma rápida el número de hits únicos de un archivo en outfmt 6, basta con aplicar una línea de código bash en la que se corta la primera columna del archivo tabular, se ordena, se eliminan los redundantes y

se cuentan los ítems:

```
cut -f 1 tab_blastxresult_allgprod0007291_c500iso | sort | uniq |  
wc
```

Un ejemplo de resultado que se obtiene es:

```
1132 1132 13584
```

Es decir, 1.132 isotigs (1.132 líneas, 1.132 palabras, 1.3584 letras) se han obtenido como hits únicos del determinado blast.

Sin embargo, normalmente, a lo que se aspira es al archivo de hits de blast en formato fasta, y esto no lo proporciona ninguna de las opciones de blast.

Para eso diseñamos un script que ayudó al análisis de resultados de los múltiples blast que realizamos ahorrando muchísimo tiempo y de forma muy eficaz:

```
#!/usr/bin/perl  
  
# Extract specific sequences from a fasta file (as  
# first argument) according to a list  
# of gene names in a .txt file (as second argument)  
(*NOTE:gene names may be redundant),  
# exporting three output files:  
# - File 1: fasta file  
# - File 2: gene name - nucleotide sequence (as  
# third argument)  
# - File 3: nucleotide sequences (as fourth  
# argument).
```

```

# Usage example
# perl get_specific_seqs fasta gene_names_file
out1 out2 out3

use warnings;
use strict;

# Open fasta file and text file from arguments
print "Need three filenames as arguments!\n"
and exit if @ARGV != 5;
open( my $FH, '<', $ARGV[0]
      or die "Couldn't open file
\"$ARGV[0]\":!\n";
open( my $FH_b, '<', $ARGV[1]
      or die "Couldn't open file
\"$ARGV[1]\":!\n";

# Save output files
open( my $out_FH, '>', $ARGV[2]
      or die "Couldn't save file \"$ARGV[2]\":!\n";
open( my $out2_FH, '>', $ARGV[3]
      or die "Couldn't save file \"$ARGV[3]\":!\n";
open( my $out3_FH, '>', $ARGV[4]
      or die "Couldn't save file \"$ARGV[4]\":!\n";

# Store headers as keys and sequences as values
of the first fasta file in a new hash '%f'
my %f;
my $header;
while( my $line = <$FH> ) {
    chomp $line;
    if( $line =~ m/^(\\w+\\d+) .+$/ ) { # Note:
headers should maintain only the first word and
not the '>' sign
        $header = $1;
    }
}

```

```

else {
    next if !defined $header;
    $line =~ s/\s//g;
    ${$header} .= $line;
}
}
close $FH;

# Store gene names of the .txt file in an array '@a'
my @a = <$FH_b>;
chomp @a;
#print "@a", "\n"; exit;

# Print those sequences of the fasta file containing
the gene names of the .txt file, in the
# same order of the .txt file.
my @b;
foreach my $i ( @a ) {
    if( exists ${$i} ) {
        print {$out_FH} '>', $i, "\n", ${$i}, "\n";
# prints fasta
        print $i, "\n", ${$i}, "\n";          # prints
gene name - sequence to the screen
        print {$out2_FH} $i, "\n", ${$i}, "\n";
# prints gene name - sequence to output file 1
        print {$out3_FH} ${$i}, "\n";      #
prints sequences to output file 2
        push ( @b, ${$i} );
        # stores sequences in an array to
check its number afterwards
    }
}

print "Number of sequences in original fasta file:
", scalar keys %f, "\n";
print "Number of selected sequences: ", scalar @b,
"\n";

```

Por supuesto, también es importante un método rápido de extracción de otra información relevante de las secuencias. Para ello se pueden diseñar scripts para extraer la información deseada (nombre de la secuencia, e-value, longitud de la secuencia, etc.) de forma rápida y sistemática, para realizar las estadísticas que pudieren interesar. Esto es especialmente fácil desde el archivo de salida en formato tabular, ya que sólo tienen que indicarse las columnas de interés para colocarlas en un archivo nuevo (y esto son simples órdenes de bash).

Desde un punto de vista global, el proceso de búsqueda de genes basado en la similitud de secuencia puede presentar diversas complicaciones, especialmente en cuanto al formato en que pudieren encontrarse las secuencias o transcriptomas que nos interesan.

Nosotros nos encontramos con un ligero inconveniente, por ejemplo, en el momento concreto de obtener las secuencias de *Tribolium castaneum* del trabajo de Prince et al. (2010): los autores no nos proporcionaron las secuencias específicas de su trabajo, sino solo el conjunto de ellas en un archivo fasta. Esto nos obligó a diseñar un script específico para obtener sólo aquellas de interés del transcriptoma, i.e., los genes de expresión masculina identificados por su estudio:

```
#!/usr/bin/perl

# Extract specific sequences from a fasta file (as first argument)
according to a list
# of gene names in a .txt file (as second arguments), exporting the
results in a new
# file (as third argument).

# Usage example
# perl get_specific_seqs fasta gene_names_file output_file

use warnings;
use strict;

# Open fasta file and text file from arguments
print "Need three filenames as arguments!\n" and exit if @ARGV
!= 3;
```

```

open( my $FH, '<', $ARGV[0]
  or die "Couldn't open file \"\$ARGV[0]\":!\n";
open( my $FH_b, '<', $ARGV[1]
  or die "Couldn't open file \"\$ARGV[1]\":!\n";

# Save output file
open( my $out_FH, '>', $ARGV[2]
  or die "Couldn't save file \"\$ARGV[2]\":!\n";

# Store headers as keys and sequences as values of the first fasta file
in a new hash '%f'
my %f;
my $header;
while( my $line = <$FH> ) {
  chomp $line;
  if( $line =~ m/^(>.+)$/ ) { # Note: headers should maintain the '>'
sign
    $header = $1;
  }
  else {
    next if !defined $header;
    $line =~ s/\s//g;
    ${f{$header}} .= $line;
  }
}
close $FH;

# Store gene names of the .txt file in an array '@a'
my @a = <$FH_b>;
chomp @a;

# Extract those sequences of the fasta file containing the gene names
of the .txt file
# in a new hash '%g'
my %g;
foreach my $k ( keys %f ) {
  foreach my $i ( @a ) {
    if( $k =~ m/$i/ ) {
      $g{$k} = $f{$k};
    }
  }
}

```



```
    }
    #print keys %g, "\n", values %g, "\n";
    #print "@{[%g]}";
    # Print to the screen and save in the new file
    foreach my $k ( keys %g ) {
        print $k, "\n", $g{$k}, "\n";
        print {$out_FH} $k, "\n", $g{$k}, "\n";
    }
    print "Number of sequences in original fasta file: ", scalar keys %f,
    "\n";
    print "Number of selected sequences: ", scalar keys %g, "\n";
```

Así, como se observa, las opciones de programación que ofrece Perl son vastas, tan vastas como *El Lenguaje*, y tan limitado como él.

### 11.3.3. Gene-finding en la base de datos de transcriptomas de insectos 1KITE

1KITE es un proyecto internacional que aspira a la secuenciación de 1000 transcriptomas de insectos. En nuestra colaboración con el ZFMK, utilizamos 19 transcriptomas de 1KITE, todos ellos de especies de coleópteros.

La primera aproximación a la búsqueda de genes en los transcriptomas de 1KITE, realizada mediante algoritmos de blast, generó un gran número de hits (mínimo: 256 en *Hydrochara caraboides*; máximo: 1.104 en *Lepicerus sp.*), que extrajimos con nuestro script de extracción de secuencias de resultados de blast. Como apunte, diremos que en este paso tuvimos que modificar la *regular expression* del script de extracción de secuencias, ya que los encabezamientos de las secuencias de los fastas de 1KITE diferían en código a los de *Calligrapha sp.*:

```
if( $line =~ m/^(.+)$/ )
```

Sin embargo, inmediatamente después, se optó por sustituir la búsqueda de genes mediante blast por realizarlo mediante Orthograph (Petersen et al. 2017), cuando se observó que la calidad de las secuencias de los hits de éste era muy superior.

Este programa, utilizando únicamente tres fuentes de datos (los genes diana (a buscar) codificados según la base de ortólogos OrthoDB, los transcriptomas (a escanear) y tres genomas de referencia —mínimo tres y típicamente de especies modelo; en nuestro caso: *D. melanogaster*, *T. castaneum* y *Acromyrmex echinatio*—), es capaz de rescatar aquellos transcritos que por cuya similitud de secuencia corresponderían a los putativos homólogos de los genes diana.

La potencia de Orthograph se basa en la construcción previa de un alineamiento de los ortólogos del gen de interés en tres (o más) especies de referencia (suficientemente distantes), que es capaz de atraer con eficacia los transcritos de un transcriptoma con enorme similitud de secuencia a dichos ortólogos de referencia.

El éxito se debe en parte a la calidad de la base grupos de ortólogos OrthoDB (Kriventseva et al. 2015), estrictamente curada. Ciertamente, viene a colación que dicha base de ortólogos se actualizó durante el proceso de análisis, pasando de la

versión v8 a la 9.1 (Zdobnov et al. 2017) obligándonos a revisar todos los grupos de ortología y sus códigos de identificación, utilizados en los scripts, el resultado de lo cual fue satisfactorio al comprobar que gran parte de la actualización era conservativa.

Otro de los éxitos de obtención de hits de Orthograph es la utilización de blastx como esqueleto central de sus búsquedas por similitud. La utilización de una base de referencia proteica facilita la anotación de transcritos *de novo*, ya que contrasta sus seis posibles pautas de lectura de las secuencias nuevas para atraer aquella que concuerde con un alineamiento de referencia robustamente creado, exhibiendo la función proteica conservada en sus aminoácidos.

Otra de las ventajas es que ofrece los resultados de la búsqueda por similitud tanto en nucleótidos como en aminoácidos, y en directorios inteligentemente ordenados.

Respecto al uso del programa, este requiere la preparación de ciertos archivos previos al análisis, y la adecuación de varios scripts y archivos de texto para su implementación. Éstos, pudiendo realizarse manualmente si el número de genes de búsqueda es bajo, fueron llevados a cabo mediante scripts diseñados específicamente, que agilizaron el proceso, para este y futuros análisis.

Aún así, no es un programa preparado para introducir los archivos de datos (archivo de genes a buscar, archivos de genomas de referencia y archivos de transcriptomas a peinar), sino que ciertos pasos requieren una edición y control manuales, y un necesario conocimiento del lenguaje de programación Perl para poder llevarlos a cabo.

## 11.4. Análisis filogenético

### 11.5.1. Análisis filogenético de los genes de función masculina CG9313, Tektin-A y tomboy40 en Insecta

Cuando realizamos análisis filogenéticos sobre los genes CG9313, *Tektin-A* y *tomboy40*, utilizamos conjuntamente las secuencias extraídas de los transcriptomas de *Calligrapha*, así como secuencias procedentes de bases de datos públicas como la *nr* de Genbank y genes de Flybase (Attril et al. 2016).

Sin embargo, este movimiento no fue el fundamental de la tesis, sino que funcionó a modo de una primera introspección de los datos y de las herramientas que se pudieren utilizar en el análisis mayor, a gran escala y definitivo.

Gracias a esta exploración, seguidamente decidimos analizar la evolución del conjunto de genes de individualización de espermatozoides según los acotaba Gene Ontology (GO:0007291), utilizamos Flybase (Attril et al. 2016) como referencia para la expresión génica sesgada por sexos, y orientamos los análisis a rango de Insecta y Coleoptera.

### 11.5.2. Análisis filogenético de genes de individualización de espermatozoides (GO:0007291) en Insecta y Coleoptera

Es de concordancia general que los estudios filogenéticos en Insecta no alcanzan el grado de robustez de otras clases filogenéticas, debido a la gran diversidad de especies que contiene, que deriva en importantes lagunas de representación, por la ausencia de secuenciación de sus genomas o transcriptomas. Y el mismo problema es heredado en Coleoptera, el Orden mayoritario de Insecta, ya que los coleópteros suponen el ~40% total de las especies de insectos y el ~25% de las especies descritas del planeta (McKenna et al. 2015) y que 90% de la diversidad de coleópteros está aún por categorizar (Benton et al. 2016).

Para lidiar con este problema durante el análisis filogenético en Insecta, las secuencias de coleópteros del proyecto 1KITE y de nuestro laboratorio han sido claves para compensar las lagunas presentes en OrthoDB.

Uno de los retos logísticos de este trabajo consiste en analizar un número considerable de genes (44) simultáneamente. Aunque en un principio utilizamos jModelTest2 para encontrar el modelo evolutivo más adecuado para conjunto de datos, los estudios definitivos se realizaron con SMS y PhyML (Lefort et al. 2017), obteniendo agilidad en el proceso de análisis de múltiples datos filogenéticos simultáneamente. Asimismo, varias cuentas en el portal CIPRES (Miller et al. 2010) ejecutadas de forma paralela ayudaron firmemente a agilizar la construcción de árboles de los 44 genes. El trabajo con Tracer (Rambaut et al. 2014) y FigTree (2016) para múltiples genes puede ser tedioso pero no complicado.

Uno de los objetivos interesantes del análisis filogenético es que nos permitió identificar rápidamente genes de copia múltiple, descubrir putativos parálogos y separarlos de los putativos ortólogos; un paso clave para proseguir el análisis subsiguiente en materia de evolución molecular de genes.

### **11.5. Evolución molecular de genes de individualización de espermatozoides (GO:0007291) en Insecta y Coleoptera**

De la literatura puede desprenderse la idea de que existe cierta controversia entre aquellos que defienden que los genes de función masculina (*male-function genes*) presentan tasas evolutivas más altas (Wyckoff et al. 2000; Swanson et al. 2001; Swanson & Vacquier 2002; Civetta 2003; Gao et al. 2003; Swanson et al. 2003; Zhang et al. 2004; Inoue et al. 2005; Clark & Dell 2006; Dorus et al. 2006; Gasper & Swanson 2006; Panhuis et al. 2006; Ellegren & Parsch 2007; Mank et al. 2007; Proschel et al. 2006; Singh et al. 2009; Krzywinska et al. 2009; Mank&Ellegren, 2009; Almeida et al. 2009; Dorus et al. 2010; Magnusson et al. 2011; Azevedo et al, 2012; Rettie&Dorus, 2013) y aquellos que no detectan tales diferencias (Dorus et al. 2006; Dorus et al. 2011; Rettie & Dorus, 2012; Singh and Jagadeeshan, 2013) (Tabla 1).

**Tabla 4.** Resumen de genes en los que se ha detectado respectivamente selección positiva o selección purificadora

Selección positiva	Selección purificadora	Referencia
<i>Pkd2</i>		Gao et al, 2003
accessory gland genes		Dorus et al, 2006
ACP genes		Dorus et al, 2006
Sex-biased loci		Zhang et al, 2004; Pröschel et al, 2006; Ellegren & Parsch 2007; Mank et al, 2007; Mank&Ellegren, 2009
<i>Zonadhesin</i>		Swanson and Vacquier 2002; Civetta 2003; Swanson et al, 2003; Inoue et al, 2005; Gasper & Swanson 2006; Dorus et al, 2010
<i>Zona pellucida 3 receptor</i>		Swanson and Vacquier 2002; Civetta 2003; Swanson et al, 2003; Inoue et al, 2005; Gasper & Swanson 2006; Dorus et al, 2010
<i>Izumo1</i>		Swanson and Vacquier 2002; Civetta 2003; Swanson et al, 2003; Inoue et al, 2005; Gasper & Swanson 2006; Dorus et al, 2010
<i>Adam</i> gene family		Swanson and Vacquier 2002; Civetta 2003; Swanson et al, 2003; Inoue et al, 2005; Gasper & Swanson 2006; Dorus et al, 2010
membrane and acrosomal sperm proteins		Dorus et al, 2010
sperm cell membrane genes (acrosomal and sperm cell surface proteins)		Dorus et al, 2010
genes involved in female and male reproduction across vertebrate and invertebrate taxa		Wyckoff et al, 2000; Swanson, Yang et al, 2001; Swanson & Vacquier 2002; Clark & Dell 2006; Dorus et al, 2010
adult male-biased <i>A. gambiae</i> genes		Magnusson et al, 2011
sperm proteins (such as binding factors)		Singh & Jagadeeshan, 2012
8% of the sperm genes		Rettie&Dorus, 2013
seminal fluids proteins		
	DmSP	Dorus et al, 2006
	overall sperm proteome	Dorus et al, 2006; Singh et al, 2009
	sperm-egg interacting proteins	Dorus et al, 2006; Singh et al, 2009
	S-LAP	Dorus et al, 2011
	sperm proteome	Rettie&Dorus, 2012

En nuestro estudio, analizamos los 44 genes de funciones masculinas y encontramos evidencia estadística significativa para los genes sexualmente sesgados (*blanks*, *Cyt-c-d*, *gudu*, *klhl10*, *nsr*, *Prosalpha6T* y *scat*). Somos conscientes de que el alcance de nuestro análisis de evolución molecular está limitado a 44 genes y a una sola función masculina (delimitada por los criterios de Gene Ontology). Sin embargo, el número es suficiente para un estudio de significancia estadística (donde por norma general se acepta  $N \geq 30$  de tamaño muestral). Por

otro lado, la variedad de funciones descritas entre el conjunto de genes y para gen en particular también brillan por su abundancia.

Tal vez el aspecto más destacable es, que el análisis minucioso de la evolución molecular de cada uno de los 44 genes de forma individualizada, detallando los eventos de duplicación (putativos parálogos) y las putativas pérdidas de genes en cada linaje genético, supone un pionero avance para Insecta.

Puede considerarse discutible la utilización de Flybase (Attril et al. 2016) como referencia durante el proceso de criba de parálogos y la identificación de ortólogos. Sin embargo, el cuidado preciso de dicha base de datos es ampliamente reconocido y, reiteramos ésta es la única base de estas características actualmente disponible.

Las pérdidas génicas, en cambio, son más discutibles, ya que la ausencia de la secuencia podría deberse alternativamente a razones de *missing data*, ocasionada por múltiples motivos ocurridos durante proceso experimental antes de reportar la expresión de una secuencia *de novo*. Sin embargo, esta cuestión solo puede ser resuelta cuando se repitan o completen dichos estudios genómicos.

Para estudios de nofuncionalización, neofuncionalización y subfuncionalización, esta tesis ofrece una vasta serie de datos génicos para futuros estudios de evolución molecular de estos genes 44 de forma particular, y para estudios moleculares a gran escala, ya sea en Coleoptera, Insecta o cualquier categoría superior.

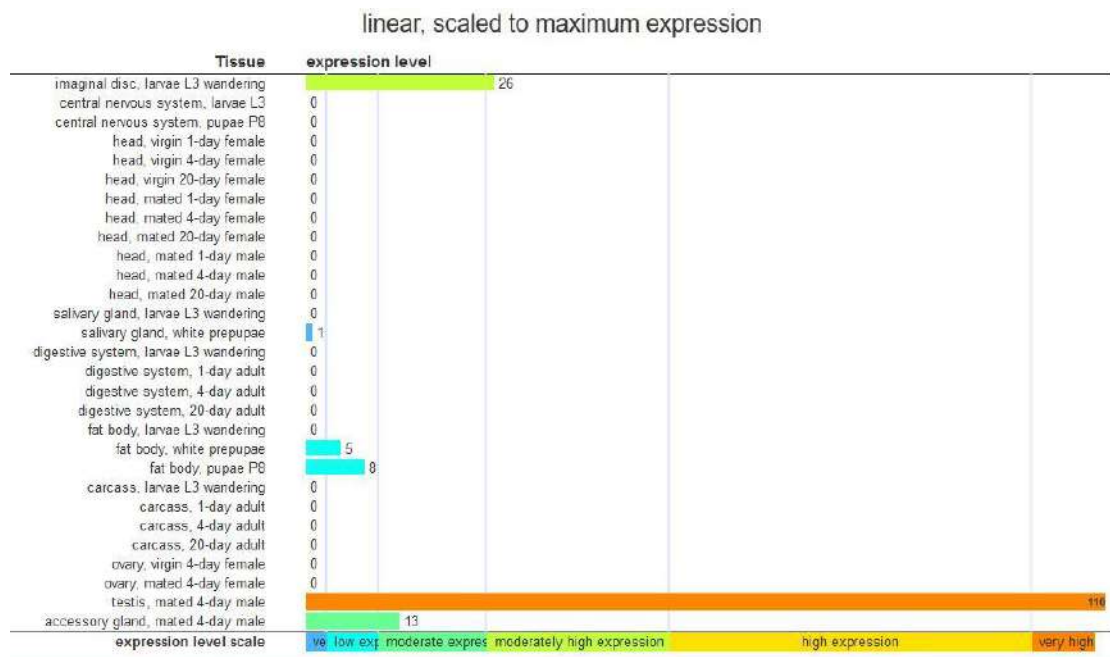


### 11.6. Redes de interacción génica de genes de individualización de espermatozoides (GO:0007291) en Insecta y Coleoptera

Existe una larga serie de puntos clave de interacción durante el proceso de expresión de un gen, desde que se encuentra codificado en la secuencia nucleotídica de DNA, hasta que su expresión tiene consecuencias para la célula, tejido, individuo o población.

Nuestro estudio nos ha permitido discernir parte del complejo sistema que rodea las funciones de un gen, desde aquéllas más directas o evidentes, a aquellas minoritarias. Analizando la expresión de RNA por tejidos y estadios del desarrollo en *Drosophila*, hemos hallado que la regla general es la expresión génica múltiple. Incluso en los genes expresados de forma sesgada en machos, encontramos variedad tisular: además de testículo, es frecuente su coexpresión en glándulas accesorias masculinas (lo cual puede considerarse esperable), en el cuerpo graso (observado en todos los casos), en las glándulas salivares (con alta frecuencia), sistema digestivo, sistema nervioso central, disco imaginal del estadio larval (donde la identificación sexual no se realiza), entre otros.

**Fig. 5.** Perfil de expresión transcriptómica por tejidos del gen *gudu* en *Drosophila*, según los informes acumulados en el proyecto modENCODE de la base de datos de Flybase (Attril et al. 2016)



Guide to modENCODE  
expression level colors

no/extremely low expression (0 - 0)
very low expression (1 - 3)
low expression (4 - 10)
moderate expression (11 - 25)
moderately high expression (26 - 50)
high expression (51 - 100)
very high expression (101 - 1000)
extremely high expression (>1000)

Por otro lado, es importante anotar que ciertos genes de función masculina (en concreto, de individualización de espermatozoides) muestran máximos de expresión en el ovario de las hembras, como *scat*, condición que no lo elimina de estar presente en la categoría de Gene Ontology de *sperm individualization* (GO:0007291).

En general se asume que la expresión de un gen en un tejido indica, por fuerza, que realiza una función allí; si bien, esto podría ser discutible: pues un gen se expresa cuando es transcrito por señales previas, lo cual no implica una acción con consecuencias futuras. Sin embargo, esta «expresión ciega», si bien, posible, no debe de ser la norma. Más razonable es afirmar que nuestro conocimiento de las funciones aún dista de ser fina y precisa. Recordemos que la función de un gen depende del contexto; este determinará la cantidad y la cualidad de las interacciones que puedan tener lugar, dando lugar a diversos escenarios posibles. El discernimiento de la función/es de un gen no es, por tanto, un proceso sencillo, y requiere tener en consideración implícita el factor ambiental.

Uno de los inconvenientes para la asignación de funciones a un gen es trasladar la consecuencia de una interacción química a un plano fenotípico superior. En nuestro caso, cabe preguntarse qué consecuencias tendría para la reproducción de un individuo, la expresión o no de cada uno de los genes de individualización de espermatozoides: ¿podría reproducirse? ¿en condiciones más o menos ventajosas? ¿qué relación hay entre la expresión de un gen y su eficacia biológica (*fitness*)?

Para contestar a todas estas preguntas, por supuesto, es necesario un estudio mucho más completo; nosotros adelantamos algunas cuestiones analizando las relaciones de un conjunto de genes de una función biológica. Encontrar todas interacciones a nivel molecular entre los productos génicos es de crucial importancia para discernir el efecto de la expresión de cualquier gen.

En nuestro estudio, además de reconstruir el mapa de interacción génica entre productos génicos que participan (con funciones moleculares diferentes) en una

misma función biológica, realizamos análisis para inferir el efecto de que grupos de genes tuvieran influencia o no en las tasas de evolución molecular de los mismos. Aunque podemos encontrarnos con un especial *missing data* en este punto (ya que la expresión génica aún dista de ser completa), nuestro análisis es un punto de partida para futuros estudios cuantitativos y cualitativos de expresión génica en Coleoptera y en Insecta y, en particular, de expresión sesgada entre sexos. Consideramos especialmente importante potenciar el estudio de expresión génica en especies más allá de *Drosophila*.

En ocasiones, se alcanza la paradoja de que cuando más se pretende abarcar adentrándose en el microcosmos de la biología para resolver una cuestión, más parece que se pierde concreción en la respuesta; esto es, parece como si la famosa ley cuadrático-cúbica de enunciada por Galileo Galilei en 1638 en su libro Dos nuevas ciencias: Discorsi e Dimostrazioni Matematiche, intorno a due nuove scienze (Fig. 5) tuviera lugar aquí, substituyéndose la variable  $v_x$  por “Nivel de abarque” y  $A_x$  por “Cantidad de conocimiento”.

Cuando un objeto se somete a un aumento proporcional en tamaño, su nuevo volumen es proporcional al cubo del multiplicador y su nueva superficie es proporcional al cuadrado del multiplicador.

Matemáticamente:

$$v_2 = v_1 \left( \frac{l_2}{l_1} \right)^3$$

donde  $v_1$  es el volumen original,  $v_2$  es el nuevo volumen,  $l_1$  es la longitud original y  $l_2$  es la nueva longitud, y:

$$A_2 = A_1 \left( \frac{l_2}{l_1} \right)^2$$

donde  $A_1$  es el área original y  $A_2$  es la nueva área.

Fig. 6. Ley cuadrático-cúbica, enunciada por Galileo Galilei en Dos nuevas ciencias: Discorsi e Dimostrazioni Matematiche, intorno a due nuove scienze (1638)

Así, el conocimiento aumentaría en un índice inferior al del nivel de abarque, y existiría, por tanto, un límite a partir del cual, cuanto más comprensivo quiera ser un estudio, más inciertas han de ser sus respuestas.

## **12. Conclusiones**

1. *De novo* annotation of five testis transcriptomes of four species of Chrysomelidae (*Calligrapha confluens*, *C. aff. floridana*, *C. multipunctata* and two specimens of *C. philadelphica*) has increased the global genetic pool of species in public databases, enlarging the genetic representativeness of coleopteran species in it, and it has provided detailed information regarding the functionality of the genes expressed in testes in these species.
2. We functionally annotated 32,8-44,6% of assembled contigs (respectively, 11.409, 10.490, 10.423, 11.656 and 11.038 isotigs of *C. confluens*, *C. floridana*, *C. multipunctata*, *C. philadelphica* (PA) and *C. philadelphica* (QC), from which 0.72–1.08% resulted functional candidates linked to reproduction or reproductive processes. The five transcriptomes were similar regarding contig richness, sequence similarity and assigned GOs in our *inter*- and *intra*- comparisons.
3. We identified 77 homologues in *Calligrapha* transcriptomes of male-biased genes occurring in *Drosophila melanogaster* and *Tribolium castaneum*, putative candidates of male-biased expression, from which we confirmed the orthologs of *CG9313*, *Tektin-A* and *tomboy40*.
4. We accomplished a detailed analysis of the molecular evolution of 44 sperm individualization genes (GO:0007291) across the Insecta class, enriching the representativeness of Coleoptera through a previous step of gene-finding in 1KITE beetle transcriptomes: *Act5C*, *Ance*, *aux*, *blanks*, *Bug22*, *CdsA*, *Chc*, *ctp*, *Cul3*, *Cyt-c-d*, *Dark*, *diddum*, *Dredd*, *Dronc*, *Duba*, *EcR*, *eIF3m*, *Dadd*, *gish*, *gudu*, *heph*, *hmw*, *jar*, *klhl10*, *Lasp*, *Mer*, *mlt*, *nes*, *Npc1a*, *nsr*, *orb2*, *Osbp*, *oys*, *Past1*, *Pen*, *poe*, *porin* *Prosalpha6T*, *scat*, *shi*, *skap*, *sw*, *Taz*, *Vps28*. Most of these genes were unbiased, but seven were male-biased (*blanks*, *Cyt-c-d*, *gudu*, *hmw*, *klhl10*, *nsr* and *Prosalpha6T*) and one female-biased (*scat*).
5. To perform gene-finding in beetles, we found Orthograph prediction pipeline —using *Drosophila melanogaster*, *Tribolium castaneum* and *Acromyrmex echinator* as OGSs— much more suitable than retrieving sequences through independent blast searches.
6. We found duplications of seven genes in all 119 insect species of Pterygota: *Ance/Acer*, *Bug22*, *klhl10*, *Npc1a/Npc1b*, *nsr*, *Pen/Kap- $\alpha$ 1/Kap- $\alpha$ 3* and *skap/Suchb*, and duplications of 18 in several evolutionary lineages of Pterygota: *Act5C*, *Ance/Acer*, *blanks*, *Bug22*, *ctp*, *Cyt-c-d*, *Cul3*, *Donc*, *gish*, *klhl10*, *Npc1a/Npc1b*, *nsr*, *orb2*, *Osbp*, *Past1*, *Pen/Kap- $\alpha$ 1/Kap- $\alpha$ 3*, *Prosalpha 6T*, and *skap/Suchb*. Specific lineage duplications were deduced for

- 10 genes: *nsr* in Acalypttratae (Diptera); *Acer* in Trichoptera, Lepidoptera and some Diptera; *skap* in Hymenoptera; *Dredd* in Ephemera; *orb2* and *Osbp* in some hemipterans; *Dronc*, *gish*, and *Past1* in Coleoptera; *blanks*, *Cul3*, in Diptera; *CdsA* in some nematocerans (midges and moth-flies); *Prosalpha6* in *Drosophila*.
7. We reported several cases of putative gene loss: *Npc1b* and *Pen* in Lepidoptera; *klhl10* in Diptera, *nsr* in aculeate Hymenoptera; *Dredd* in Culicidae (Diptera); *Duba* in Trichoptera and Lepidoptera; *Fadd* in some Hemiptera; *hmv* in *Ephemera* (Ephemeroptera) and *Anopheles* (Diptera).
  8. We inferred gene trees for each of the 41 sperm individualisation genes (*Act5C*, *Cyt-c-d* and *ctp* excluded) across the class Insecta and estimated their amino acid evolutionary rates: in average,  $0.00239 \pm 0.003012$  subs./l./Ma, ranging from 0.000237 in *orb2* to 0.009667 in *hmv*. Coleoptera resulted monophyletic in 18 of the 41 gene trees.
  9. We estimated the nucleotide substitution rates of 31 sperm individualization genes (*aux1*, *blanks*, *Chc*, *diddum*, *Dredd*, *elF3m*, *gish*, *gudu*, *Mer*, *Npc1a*, *nsr*, *orb2*, *Pen*, *poe*, *Prosalpha6T*, *scat*, *shi*, *Taz*, *CdsA*, *Cul3*, *Dark*, *Duba*, *Fadd*, *jar*, *Lasp*, *mlt*, *nes*, *Osbp*, *Past1*, *skap* and *sw*) in beetles – using a time constraint in Coleoptera of 277.4 and 315.2 Ma (deduced from the overlapping age confidence intervals: from 180.4 [52.1–328.6] Ma (*orb2*) to 451.5 Ma [288.4–633.0] (*Lasp*)— resulting, in average,  $0.00452 \pm 0.002083$  subs./l./Ma, ranging from 0.00208 subs./l./Ma in *nes* to 0.01190 subs./l./Ma in *Cul3*.
  10. We did not find significant differences in rates of evolution in proteins of genes exhibiting duplications or belonging to the same interaction network. However, we found faster rates of evolution in proteins of sex-biased genes. Surprisingly, we did not find faster evolution rates for the nucleotide sequences of the same genes in any condition: sex-biased, presence of duplications or belonging to an interaction network.
  11. We inferred a gene interaction network resulting four groups of interacting genes (*blanks*, *Cul3*, *Fadd*, *klhl10*, *skap*; *Act5C*, *Lasp*, *Dark*, *Dredd*, *Dronc*, *Duba*, *elF3m*; *aux*, *Chc*, *ctp*, *shi*, *sw*; *diddum*, *jar*, *Mer*, *Past1*) or five groups (*blanks*, *Cul3*, *Dredd*, *Fadd*, *klhl10*, *skap*; *Act5C*, *Lasp*; *Dark*, *Dronc*, *Duba*, *elF3m*; *aux*, *Chc*, *ctp*, *shi*, *sw*; *diddum*, *jar*, *Mer*, *Past1*), by using public data of specific protein-protein physical interactions (BioGRID), as well as information of genetic interactions in metabolic pathways (Flybase).

12. We did not find significant differences in rate evolution between groups of interacting genes (amino acid or nucleotide sequences). However, we found significant differences in rate evolution in some edges separating groups: *Dronc-shi* (amino acid and nucleotide sequences), *Dronc-Dredd* (nucleotide sequences) and *Chc-Past1* (nucleotide sequences).

## 13. Bibliografía



1KITE: 1000 Insect Transcriptome Evolution (2015). [online] Available at: <http://www.1kite.org/>

Almeida, F.C. & DeSalle, R. (2009). Orthology, function and evolution of accessory gland proteins in the *Drosophila repleta* group. *Genetics* 181(1): 235–245.

Altschul, S.F., Gish, W., Miller, W., Myers, E.W. & Lipman, D.J. (1990). Basic local alignment search tool. *J. Mol. Biol.* 215(3): 403-410.

Attrill, H., Falls, K., Goodman, J.L., Millburn, G.H., Antonazzo, G., Rey, A.J. & Marygold, SJ; the FlyBase Consortium (2016). FlyBase: establishing a Gene Group resource for *Drosophila melanogaster*. *Nucleic Acids Res.*, 44(D1): D786–D792.

Azevedo, R.V.D.M., Dias, D.B.S., Bretãs, J.A.C., Mazzoni, C.J., Souza N.A., Albano, R.M., Wagner, G., Davila, A.M.R. & Peixoto, A.A. (2012). The Transcriptome of *Lutzomyia longipalpis* (Diptera: Psychodidae) Male Reproductive Organs. *PLoS ONE* 7(4): e34495. doi:10.1371/journal.pone.0034495.

Baker, M. (2012). De novo genome assembly: what every biologist should know. *Nat Methods* 9(4): 333–337. doi: 10.1038/nmeth.193.

Benton, M.A., Kenny, N.J., Conrafs, K.H., Roth, S. & Lynch, J.A. (2016). Deep, staged transcriptomic resources for the novel coleopteran models *Atrachya menestriesi* and *Callosobruchus maculatus*. *PLoS ONE* 11(12): 30167431. doi:10.1371/journal.pone.0167431.

Brown, W.J. (1945). Food-plants and distribution of the species of *Calligrapha* in Canada, with descriptions of new species (Coleoptera, Chrysomelidae). *Canadian Entomologist* 77(7): 117-133.

Chen, H., Lin, L., Xie, M., Zhang, G. & Su, W. (2014). *De novo* Sequencing, Assembly and Characterization of Antennal Transcriptome of *Anomala corpulenta* Motschulsky (Coleoptera: Rutelidae). *PLoS ONE* 9(12): e114238. doi:10.1371/journal.pone.0114238.

Chu, Y. & Corey, D.R. (2012). RNA Sequencing: Platform Selection, Experimental Design, and Data Interpretation. *Nucleic Acid Ther.* 2012 Aug; 22(4): 271–274. doi: 10.1089/nat.2012.0367.

Civetta, A. (2003). Positive selection within sperm-egg adhesion domains of

fertilin: an ADAM gene with a potential role in fertilization. *Mol Biol Evol.* 20(1): 21–29.

Conesa, A., Götz, S., Garcia-Gomez, J.M., Terol, J., Talon M. & Robles, M. (2005). Blast2GO: a universal tool for annotation, visualization and analysis in functional genomics research. *Bioinformatics.* 21(18): 3674-3676.

Conesa, A. & Götz. S. (2008). Blast2GO: A Comprehensive Suite for Functional Analysis in Plant Genomics. *International Journal of Plant Genomics.* 2008 (2008): 1-13.

Dorus, S., Busby, S.A., Gerike, U., Shabanowitz, J., Hunt, D.F. & Karr, T.L. (2006). Genomic and functional evolution of the *Drosophila melanogaster* sperm proteome. *Nat Genet.* 38(12): 1440–1445.

Dorus, S., Wasbrough, E.R., Busby, J., Wilkin, E.C. & Karr, T.L. (2010). Sperm proteomics reveals intensified selection on mouse sperm membrane and acrosome genes. *Mol Biol Evol.* 27(6): 1235-1246.

Dorus, S., Wilkin, E.C. & Karr, T.L. (2011). Expansion and functional diversification of a leucyl aminopeptidase family that encodes the major protein constituents of *Drosophila* sperm. *BMC Genomics.* 12: 177. doi: 10.1186/1471-2164-12-177.

Ellegren, H. & Parsch, J. (2007). The evolution of sex-biased genes and sex-biased gene expression. *Nature Reviews Genetics.* 8(9): 689–698.

Galilei, Galileo. (1638). *Dos nuevas ciencias: Discorsi e Dimostrazioni Matematiche, intorno a due nuove scienze.* ISBN 88-02-03457-5

Gao, Z., Ruden, D.M. & Xiangyi, L. (2003). PKD2 Cation Channel Is Required for Directional Sperm Movement and Male Fertility. *Current Biology.* 13(24): 2175-2178, ISSN 0960-9822.

Gasper, J. & Swanson, W.J. (2006). Molecular population genetics of the gene encoding the human fertilization protein zonadhesin reveals rapid adaptive evolution. *Am J Hum Genet.* 79(5): 820–830.

Gene Ontology Consortium (2000). Gene ontology: tool for the unification of biology. 2000. *Nature Genet.* 25: 25-29.

Geer, L.Y., Marchler-Bauer, A., Geer, R.C., Han, L., He, J., He, S., Liu, C., Shi, W. & Bryant, S.H. (2010). The NCBI BioSystems database. *Nucleic Acids Res.* 38(Database issue): D492-6.

Gómez-Zurita, J. (2004). Resources for a phylogenomic approach in leaf beetle (Coleoptera) systematics. In: *New Developments in the Biology of Chrysomelidae*. P. Jolivet, J.A. Santiago-Blay & M. Schmitt editors. SPB Academic Publishing, The Hague, The Netherlands, pp.19-35.

Gómez-Zurita, J. (2005). New distribution records and biogeography of *Calligrapha* species (leaf beetles), in North America (Coleoptera: Chrysomelidae, Chrysomelinae). *The Canadian Naturalist*. 119(1): 88-100.

Gómez-Zurita, J. & Galián (2005). Current knowledge on genes and genomes of phytofagous beetles (Coleoptera: Chrysomeloidea, Curculionoidea): a review. *Eur. J. Entomol.* 102(4): 577-597. ISSN 1210-5759.

Gómez-Zurita, J., Funk, D.J. & Vogler, A.P. (2006). The evolution of unisexuality in *Calligrapha* leaf beetles: molecular and ecological insights on multiple origins via interspecific hybridization. *Evolution* 60(2), pp. 328-347.

Gómez-Zurita, J. (2015). What is the Leaf Beetle *Calligrapha scalaris* (Leconte)? *Breviora*. 541: 1-19. US ISSN 0006-9698.

Götz, S., García-Gómez, J.M., Terol, J., Williams, T.D., Nagaraj, S.H., Nueda, M.J., Robles, M., Talón, M., Dopazo, J. & Conesa, A. (2008). High-throughput functional annotation and data mining with the Blast2GO suite. *Nucleic Acids Research*. 36(10): 3420-3435.

Götz, S., Arnold, R., Sebastián-León, P., Martín-Rodríguez, S., Tischler, P., Jehl, M.A., Dopazo, J., Rattei, T. & Conesa, A. (2011). B2G-FAR, a species centered GO annotation repository. *Bioinformatics*. 27 (7): 919-924.

Graveley, B.R., Brooks, A.N., Carlson, J.W., Cherbas, L., Choi, J., Davis, C.A., Dobin, A., Duff, M., Eads, B., Hansen, K.D., Landolin, J., Langton, L., Malone, J., Miller, D., Roberts, J., Sandler, J., Sturgill, D., Tang, H., van Baren, M.J., Wan, K.H., Xiao, S., Yang, L., Zhang, D., Zhang, Y., Zou, Y., Andrews, J., Brenner, S.E., Brent, M., Cherbas, P., Dudoit, S., Gingeras, T.R., Hoskins, R., Kaufman, T., Oliver, B. & Celniker, S.E. (2010). The *D. melanogaster* transcriptome: modENCODE RNA-Seq data. Available at: <<http://www.modencode.org/celniker/>>

Harbers, M. & Carninci, P. (2005). Tag-based approaches for transcriptome research and genome annotation. *Nature Methods*. 2(7): 495-502.

Holley, R., Apgar, J., Everett, G., Madison, J., Marquisee, M., Merrill, S. & Zamir, A. (1965). Structure of a Ribonucleic Acid. *Science*. 147(3664): 1462-1465.

Huntley, R.P., Sawford, T., Mutowo-Muellenet, P., Shypitsyna, A., Bonilla, C., Martin, M.J. & O'Donovan, C. (2014). The GOA database: Gene Ontology annotation updates for 2015. *Nucleic Acids Research*. 43(Database issue):D1057-63. doi: 10.1093/nar/gku1113.

Inoue, N., Ikawa, M., Isotani, A. & Okabe, M. (2005). The immunoglobulin superfamily protein Izumo is required for sperm to fuse with eggs. *Nature*. 434(7030): 234–238.

i5k-Consortium (2013). The i5K Initiative: advancing arthropod genomics for knowledge, human health, agriculture, and the environment. *J Hered*. 104(5): 595-600.

Jreborg, N., Birney, E. & Durbin, R. (1999). Comparative Analysis of Noncoding Regions of 77 Orthologous Mouse and Human Gene Pairs. *Genome Res*. 9(9): 815-824.

Krzywinska, E. & Krzywinski, J. (2009). Analysis of expression in the *Anopheles gambiae* developing testes reveals rapidly evolving lineage-specific genes in mosquitoes. *BMC Genomics*. 10: 300.

Keeling, C.I., Henderson, H., Li, M., Yuen, M., Clark, E.L., Fraser, J.D., Huber, D.P.W., Liao, N.Y., Docking, T.R., Birol, I., Chan, S.K., Taylor, G.A., Palmquist, D., Jones, S.J.M. & Bohlmann, J. (2012). Transcriptome and full-length cDNA resources for the mountain pine beetle, *Dendroctonus ponderosae* Hopkins, a major insect pest of pine forests. *Insect Biochem Mol Biol*. 42: 525-536.

King, M.C. & Wilson, A.C. (1975). Evolution at two levels in humans and chimpanzees. *Science*. 188(4184): 107–116.

Kriventseva, E.V., Tegenfeldt, F., Petty, T.J., Waterhouse, R.M., Simão, F.A., Pozdnyakov, I.A., Ioannidis, P. & Zdobnov, E.M. (2015). OrthoDB v8: update of the hierarchical catalog of orthologs and the underlying free software. *Nucleic Acids Research*. 43: D250–D256.

- Kumar, A., Congiu, L., Lindstrom, L., Piironen, S., Vidotto, M. & Grapputo, A. (2014). Sequencing, De Novo assembly and annotation of the Colorado Potato Beetle, *Leptinotarsa decemlineata*, Transcriptome. *PLoS One*. 9(1): e86012.
- Kumar, S., Stecher, G., Suleski, M. & Hedges, S.B. (2017). TimeTree: A Resource for Timelines, Timetrees, and Divergence Times. *Mol Biol Evol*. 34(7): 1812–1819. doi:10.1093/molbev/msx116.
- Lande, R. (1980). Sexual dimorphism, sexual selection, and adaptation in polygenic characters. *Evolution*. 34(2): 292-305.
- Lefort, V., Longueville, J.E. & Gascuel, O. (2017). SMS: Smart Model Selection in PhyML. *Mol. Biol. Evol*. 34(9): 2422–2424. doi: 10.1093/molbev/msx149.
- Leung, Y.Y., Ryvkin, P., Ungar, L.H., Gregory, B.D. & Wang, L.S. (2013). CoRAL: predicting non-coding RNAs from small RNA-sequencing data. *Nucleic Acids Research*. 41(14): e137. doi: 10.1093/nar/gkt426.
- Magnusson, K., Mendes, A.M., Windbichler, N., Papathanos, P.A., Nolan, T., Dottorini, T., Rizzi, E., Christophides, G.K. & Crisanti, A. (2011). Transcription regulation of sex-biased genes during ontogeny in the malaria vector *Anopheles gambiae*. *PLoS One*. 6(6): e21572. doi: 10.1371/journal.pone.0021572.
- Marguerat, S. & Bähler, J. (2010). RNA-seq: from technology to biology. *Cell. Mol. Life Sci*. 67: 569–579. doi 10.1007/s00018-009-0180-6.
- McKenna, D.D., Wild, A.L., Kanda, K., Bellamy, C.L., Beutel, R.G., Caterino, M.S., Farnum, C.W., Hawks, D.C., Ivie, M.A., Jameson, M.L., Leschen, R.A.B., Marvaldi, A.E., Mchugh, J.V., Newton, A.F., Robertson, J.A., Thayer, M.K., Whiting, M.F., Lawrence, J.F., Ślipinski, A., Maddison, D.R. & Farrell, B.D. (2015). The beetle tree of life reveals that Coleoptera survived end-Permian extinction to diversity during the Cretaceous terrestrial revolution. *Systematic Entomology*. 40(4): 835-880. doi: 10.1111/syen.12132.
- Meisel, R.P., Malone, J.H., & Clark, A.G. (2012). Disentangling the relationship between sex-biased gene expression and X- linkage. *Genome Research*. 22(7): 1255–1265.
- Meisel, R.P., Hilldorfer, B.B., Koch, J.L., Lockton, S. & Schaeffer, S.W. (2019). Adaptive evolution of genes duplicated from the *Drosophila pseudoobscura* neo-X chromosome. *Molecular Biology and Evolution*. 27(8): 1963–1978.

- Metzker, M. (2010). Sequencing technologies — the next generation. *Nat Rev Genet.* 11: 31–46. doi: [10.1038/nrg2626](https://doi.org/10.1038/nrg2626).
- Meisel, R.P., Malone, J.H. & Clark, A.G. (2012). Disentangling the relationship between sex-biased gene expression and X-linkage. *Genome Research.* 22(7): 1255–1265.
- Meisel, R.P., Hilldorfer, B.B., Koch, J.L., Lockton, S. & Schaeffer, S.W. (2019). Adaptive evolution of genes duplicated from the *Drosophila pseudoobscura* neo-X chromosome. *Molecular Biology and Evolution.* 27(8): 1963–1978.
- Miller, M.A., Pfeiffer, W. & Schwartz, T. (2010). Creating the CIPRES Science Gateway for inference of large phylogenetic trees. *Proceedings of the Gateway Computing Environments Workshop (GCE), 14 Nov. 2010, New Orleans, LA*, pp 1–8.
- Montelongo, T., Gómez-Zurita, J. (2015). Non-random patterns of genetic admixture expose the complex historical hybrid origin of unisexual leaf beetle species in the genus *Calligrapha*. *American Naturalist.* 185(1): 113–134.
- Normark (2003). The evolution of alternative genetic systems in insects. *Annu. Rev. Entomol.* 48: 397–423. doi: [10.1146/annurev.ento.48.091801.112703](https://doi.org/10.1146/annurev.ento.48.091801.112703).
- Oppenheim, S.J., Baker, R.H., Simon, S. & DeSalle, R. (2015). We can't all be supermodels: the value of comparative transcriptomics to the study of non-model insects. *Insect Molecular Biology.* 24(2): 139–154. doi: [10.1111/imb.12154](https://doi.org/10.1111/imb.12154).
- Ozsolak, F. & Milos, P.M. (2011). RNA sequencing: advances, challenges and opportunities. *Nature reviews. Genetics.* 12: 87–98.
- Panhuis, T.M., Clark, N.L., Swanson, W.J. (2006). Rapid evolution of reproductive proteins in abalone and *Drosophila*. *Philos Trans R Soc Lond B Biol Sci.* 361(1466): 261–268.
- Parisi, M., Nuttall, R., Naiman, D., Bouffard, G., Malley, J., Andrews, J., Eastman, S. & Oliver, B. (2003). Paucity of genes on the *Drosophila* X-chromosome showing male-biased expression. *Science.* 299(5607): 697–700.
- Parkinson, J. & Blaxter, M. (2009). Expressed Sequence Tags: An Overview. In:

*Expressed Sequence Tags (ESTs). Generation and Analysis. Methods in Molecular Biology* 533. Humana Press. doi:10.1007/978-1-60327-136-3\_1.

Pauchet, Y., Wilkinson, P., van Munster, M., Augustin, S., Pauron, D. & Ffrench-Constant, R.H. (2009). Pyrosequencing of the midgut transcriptome of the poplar leaf beetle *Chrysomela tremulae* reveals new gene families in Coleoptera. *Insect Biochem Mol Biol.* 39(5-6): 403-13. doi: 10.1016/j.ibmb.2009.04.001.

Petersen, M., Meusemann, K., Donath, A., Dowling, D., Liu, S., Peters, R.S., Podsiadlowski, L., Vasilikopoulos, A., Zhou, X., Misof, B. & Niehuis, O. (2017). Orthograph: a versatile tool for mapping coding nucleotide sequences to clusters of orthologous genes. *BMC Bioinformatics.* 18(1): 111. doi 10.1186/s12859-017-1529-8.

Ranz, J.M., Castillo-Davis, C.I., Meiklejohn, C.D. & Hartl, D.L. (2003). Sex-Dependent Gene Expression and Evolution of the *Drosophila* Transcriptome. *Science.* 300(5626): 1742-1745.

Rambaut, A. (2016). FigTree v1.4.3. Available at: <<http://tree.bio.ed.ac.uk/software/figtree/>>

Rambaut, A., Suchard, M.A., Xie, D. & Drummond, A.J. (2014). Tracer v1.6. Available at: <<http://beast.bio.ed.ac.uk/Tracer>>

Ruan, Y., Le Ver, P., Ng, H.H. & Liu, E.T. (2004). Interrogating the transcriptome. *Trends Biotechnol.* 22: 23–30.

Rettie, E.C. & Dorus, S. (2012). *Drosophila* sperm proteome evolution: Insights from comparative genomic approaches. *Spermatogenesis.* 2(3): 213-223. doi: 10.4161/spmg.21748. PMID: 23087838; PMCID: PMC3469443.

Scolari, F., Benoit, J., Michalkova, V., Aksoy, E., Takac, P., Abd-Alla, A.M.M., Malacrida, A.R., Aksoy, S. & Attardo, G.M. (2016). The Spermatophore in *Glossina morsitans morsitans*: Insights into Male Contributions to Reproduction. *Sci. Rep.* 6: 20334. doi: 10.1038/srep20334.

Shendure, J., Ji, H. (2008). Next-generation DNA sequencing. *Nat Biotechnol.* 26: 1135–1145. <https://doi.org/10.1038/nbt1486>.

Singh, R. & Jagadeeshan, S. (2012). Sex and Speciation: *Drosophila* Reproductive Tract Proteins—Twenty Five Years Later. *International Journal of Evolutionary*

*Biology*. Volume 2012, Article ID 191495. doi:10.1155/2012/191495.

Swanson, W.J., Clark, A.G., Waldrip-Dail, H.M., Wolfner, M.F., Aquadro, C.F. (2001). Evolutionary EST analysis identifies rapidly evolving male reproductive proteins in *Drosophila*. *Proc Natl Acad Sci*. 98(13): 7375–7379.

Swanson, W.J., Vacquier, V.D. (2002). The rapid evolution of re- productive proteins. *Nat Rev Genet*. 3(2): 137–144.

The UniProt Consortium (2017). UniProt: the universal protein knowledgebase. *Nucleic Acids Res*. 45(D1): D158-D169.

Theodorides, K., De Riva, A., Gómez-Zurita, J., Foster, P.G. & Vogler, A.P. (2002). Comparison of EST libraries from seven beetle species: towards a framework for phylogenomics of the Coleoptera. *Insect Molecular Biology*. 11(5): 467–475.

Trapnell, C., Pachter, L. & Salzberg, S.L. (2009). TopHat: discovering splice junctions with RNA-Seq, *Bioinformatics*. 5(9): 1105–1111. doi: 10.1093/bioinformatics/btp120.

Wang, Z., Gerstein, M. & Snyder, M. (2009). RNA-Seq: a revolutionary tool for transcriptomics. *Nature Reviews Genetics*. 10(1): 57–63. doi:10.1038/nrg2484.

Wyckoff, G.J., Wang, W. & Wu, C.I. (2000). Rapid evolution of male reproductive genes in the descent of man. *Nature*. 403(6767): 304–309.

Wolfner, M.F. (2002). The gifts that keep on giving: physiological functions and evolutionary dynamics of male seminal proteins in *Drosophila*. *Heredity*. 88: 85–93.

Zhao, Q., Wang, Y., Kong, Y., Luo, D., Li, X, & Hao, P. (2011). Optimizing de novo transcriptome assembly from short-read RNA-Seq data: a comparative study. *BMC bioinformatics*. 12(S2).



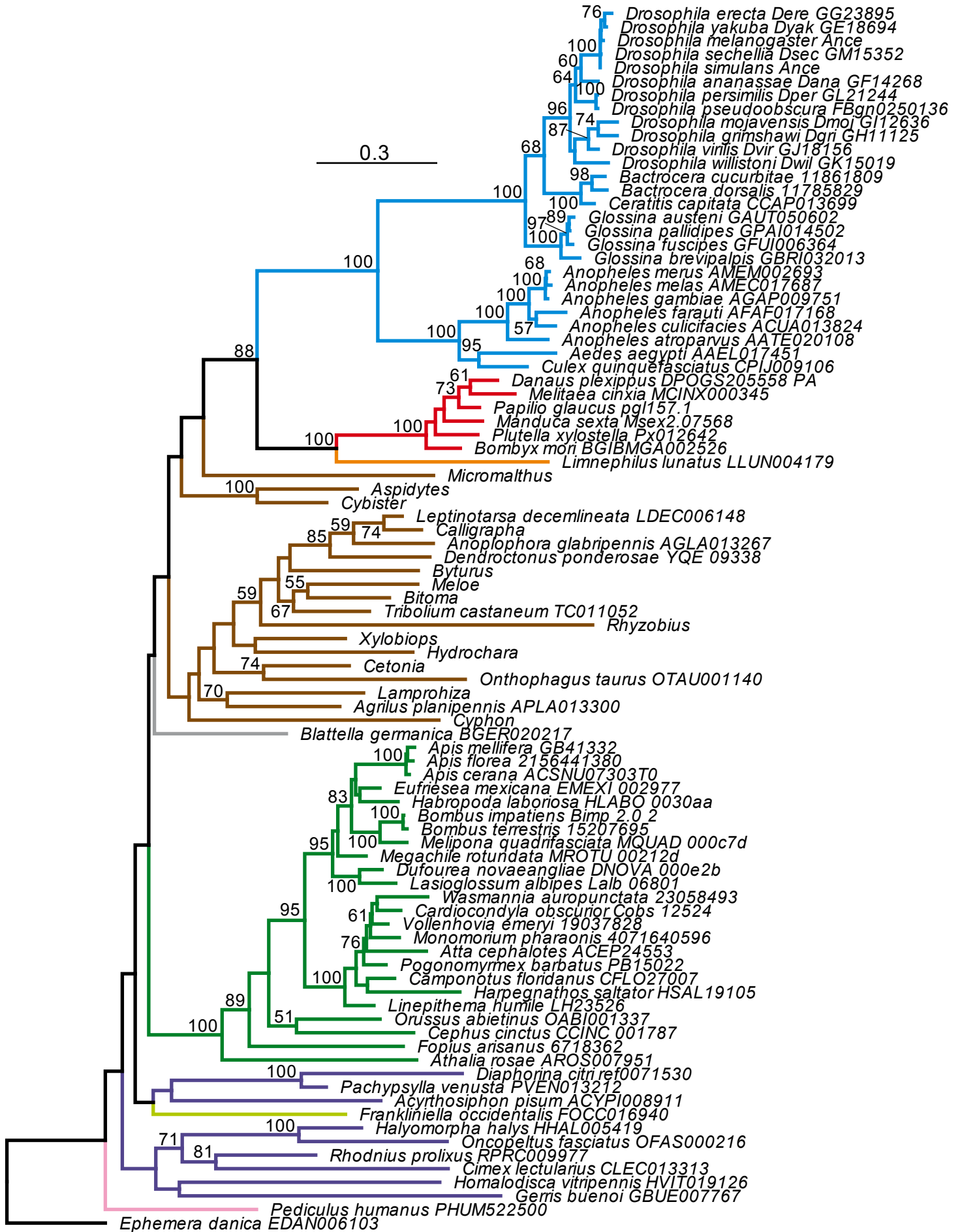
## **14. Anexo**

**File S1.** Maximum likelihood trees based on the amino acid alignments of different sperm individualization proteins in insects.

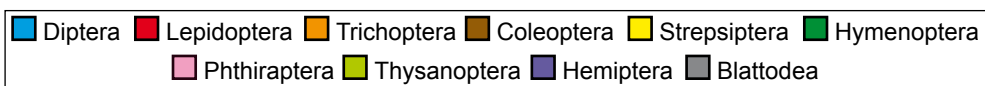
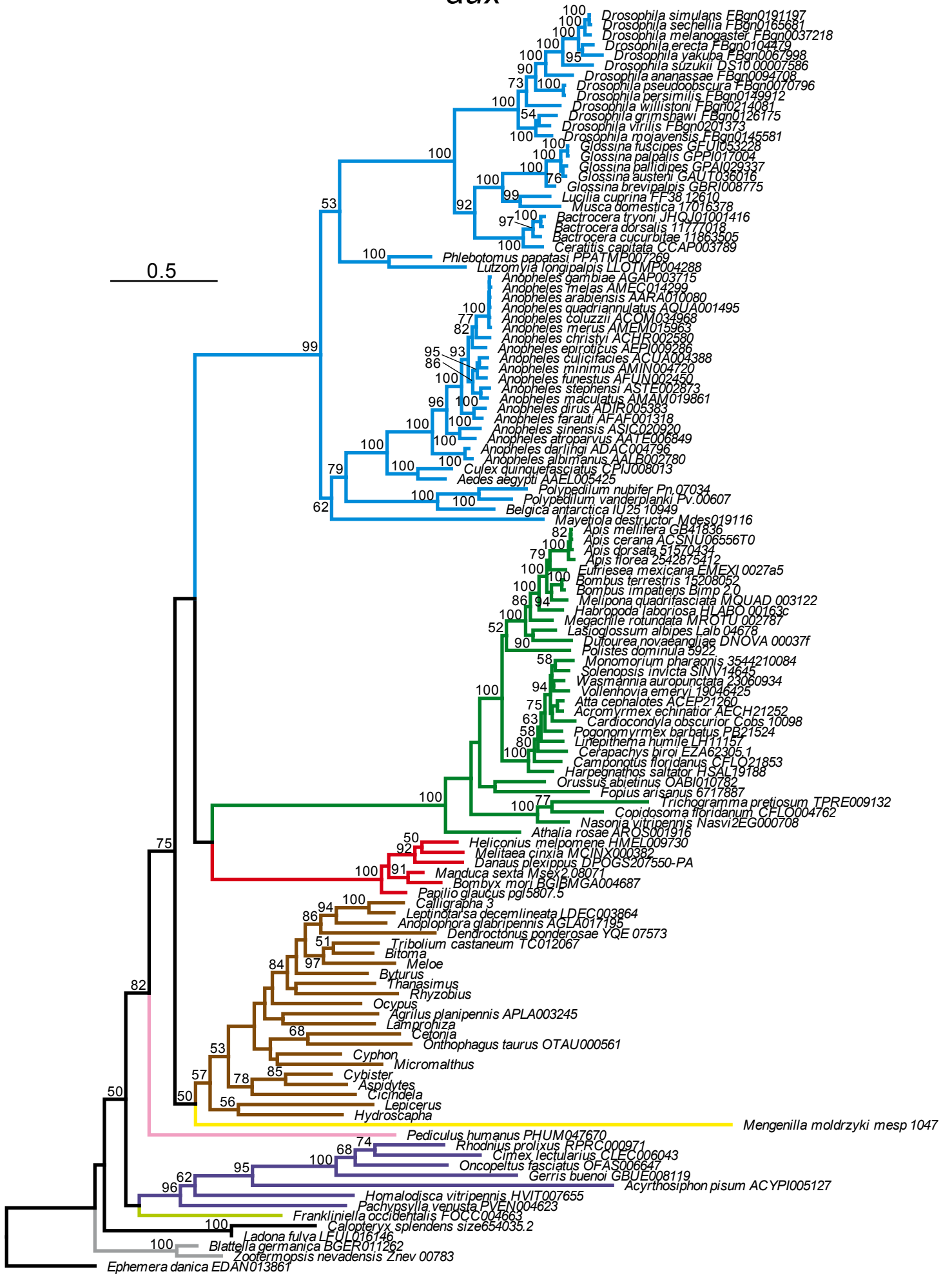
Tree	Gene	Model	Gamma	Invariant	Log likelihood
1	<i>Ance*</i>	LG	0.975	0.075	-36264.30406
2	<i>aux</i>	LG	0.861	0.042	-101341.43324
3	<i>blanks</i>	JTT	1.000	0.006	-51030.51327
4	<i>Bug22*</i>	LG	0.970	0.076	-11344.29415
5	<i>CdsA</i>	JTT	0.662	0.164	-20188.93347
6	<i>Chc</i>	JTT	0.646	0.429	-27341.73092
7	<i>Cul3*</i>	JTT	0.617	0.203	-21620.12771
8	<i>Dark</i>	JTT	1.873	0.003	-175174.99008
9	<i>didum</i>	LG	0.896	0.094	-118892.32085
10	<i>Dredd*</i>	JTT	1.420	0.006	-57705.56572
11	<i>Dronc</i>	WAG	1.405	0.013	-53433.37737
12	<i>Duba</i>	JTT	0.992	0.055	-43655.01217
13	<i>EcR</i>	JTT	0.663	0.140	-13958.07538
14	<i>eIF3m</i>	JTT	0.817	0.145	-15809.34033
15	<i>Fadd</i>	JTT	1.307	0.004	-30734.15211
16	<i>gish</i>	JTT	0.678	0.335	-7663.02559
17	<i>gudu</i>	LG	0.921	0.038	-38783.13178
18	<i>heph</i>	JTT	0.965	0.170	-15480.89741
19	<i>hmw</i>	JTT	1.149	0.007	-35247.66673
20	<i>jar</i>	LG	0.857	0.181	-62829.80003
21	<i>klhl10*</i>	LG	0.913	0.017	-35556.78931
22	<i>Lasp</i>	JTT	0.844	0.249	-9432.56476
23	<i>Mer</i>	JTT	0.699	0.123	-21840.64371
24	<i>mlt</i>	LG	0.628	0.095	-34633.21523
25	<i>nes</i>	LG	1.073	0.040	-42646.95754
26	<i>Npcla*</i>	LG	0.833	0.156	-85235.55965
27	<i>nsr*</i>	JTT	0.854	0.036	-28961.01227
28	<i>orb2</i>	JTT	0.489	0.214	-3222.50637
29	<i>Osbp</i>	JTT	0.863	0.082	-51127.87538
30	<i>oys</i>	LG	0.984	0.066	-29099.88610
31	<i>Past1</i>	LG	0.711	0.302	-16096.65909
32	<i>Pen*</i>	LG	0.930	0.090	-38161.29725
33	<i>poe</i>	JTT	0.769	0.060	-163837.44235
34	<i>porin</i>	LG	1.341	0.028	-15895.84791
35	<i>Prosalpha6T</i>	LG	0.792	0.197	-15675.74518
36	<i>scat</i>	JTT	0.939	0.037	-72098.82783
37	<i>shi</i>	LG	0.650	0.287	-23338.29850
38	<i>skap*</i>	LG	0.831	0.257	-19572.71029
39	<i>sw</i>	JTT	0.960	0.066	-27022.45452
40	<i>Taz</i>	LG	1.034	0.157	-17370.63510
41	<i>Vps28</i>	LG	0.720	0.100	-6256.72767

\*Orthologs to the *Drosophila* paralog with sperm individualization function.

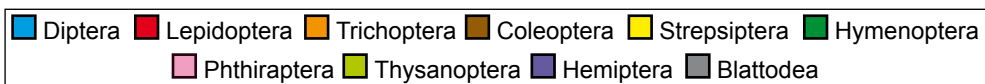
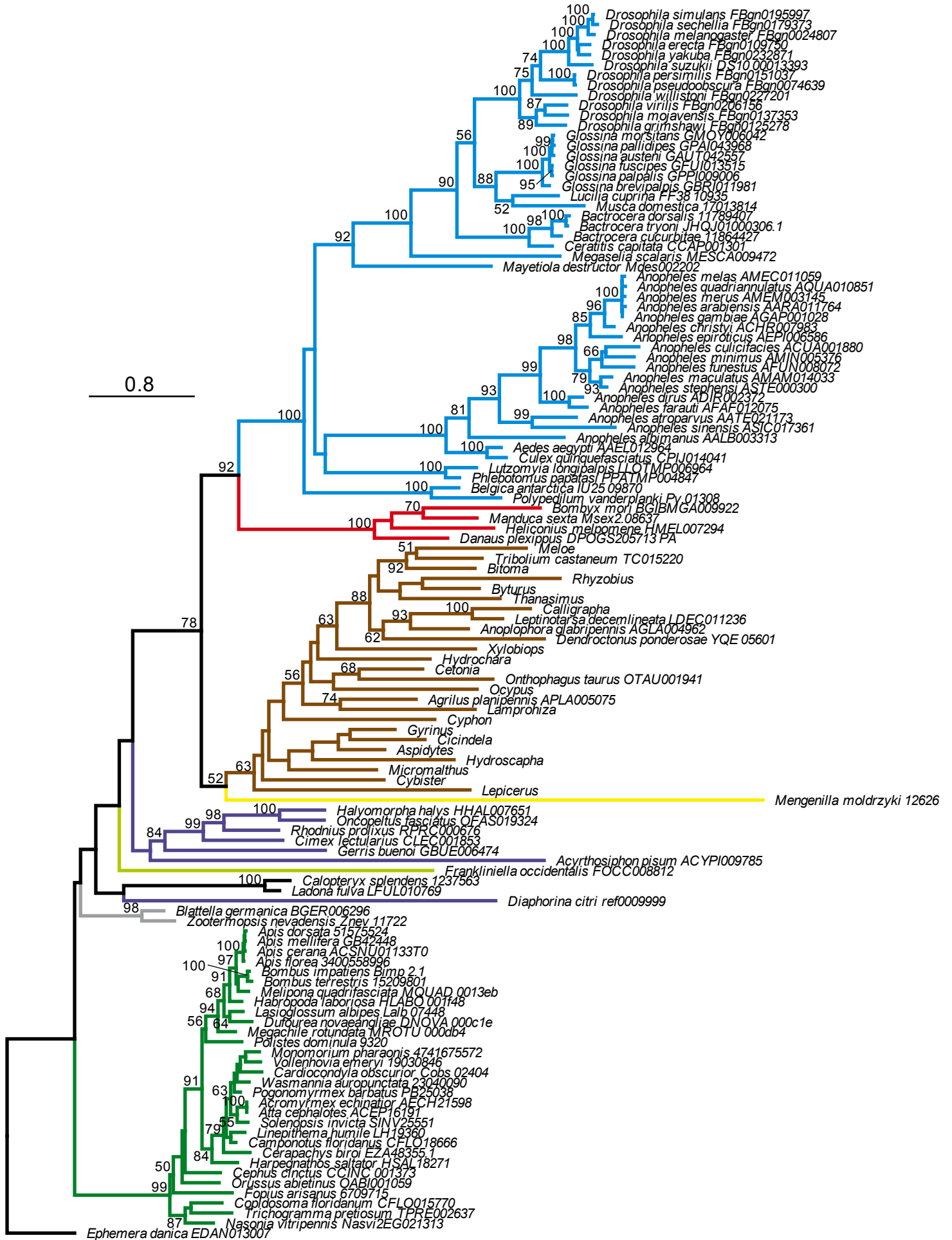
# Ance



# aux

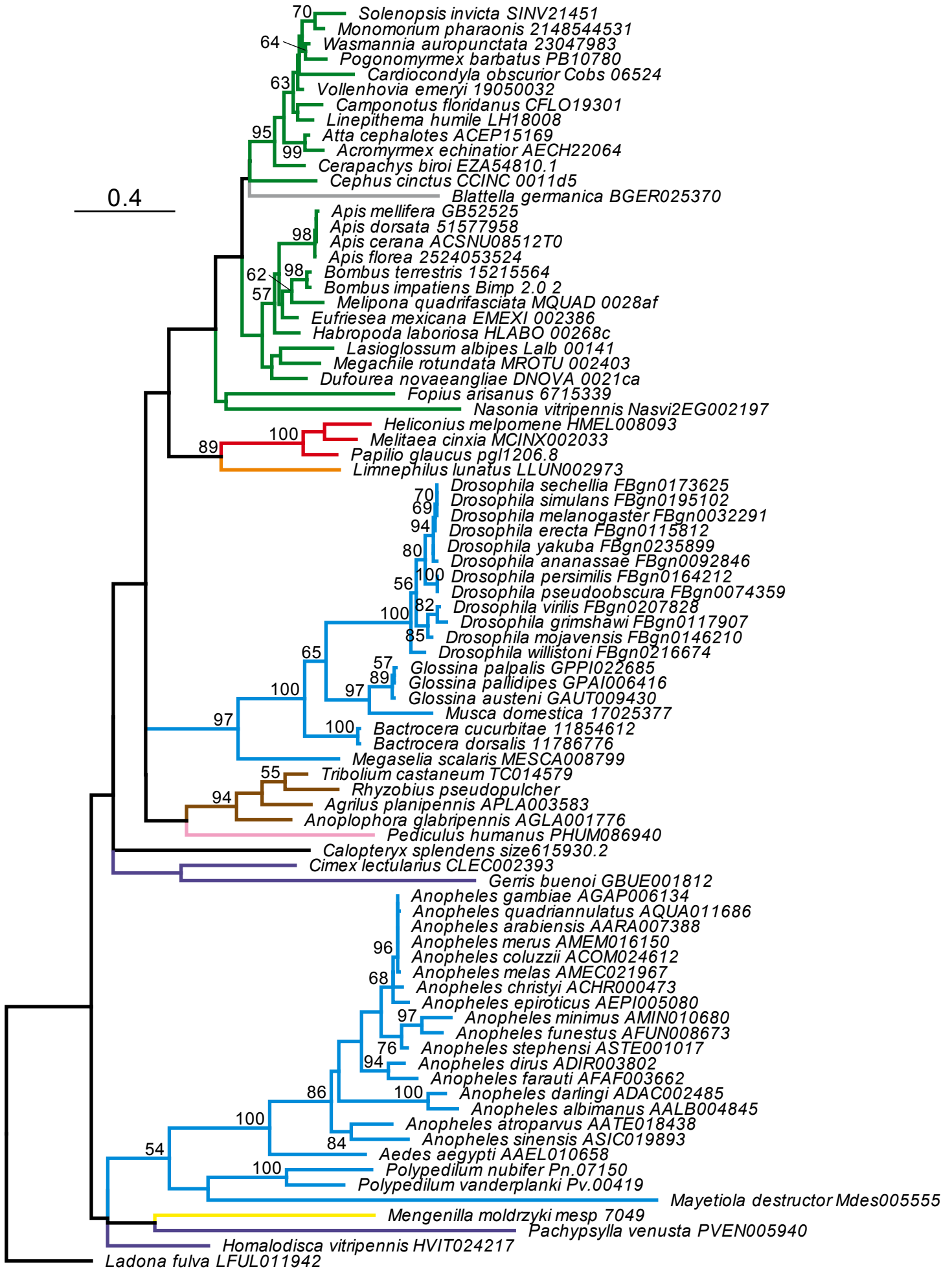


# blanks

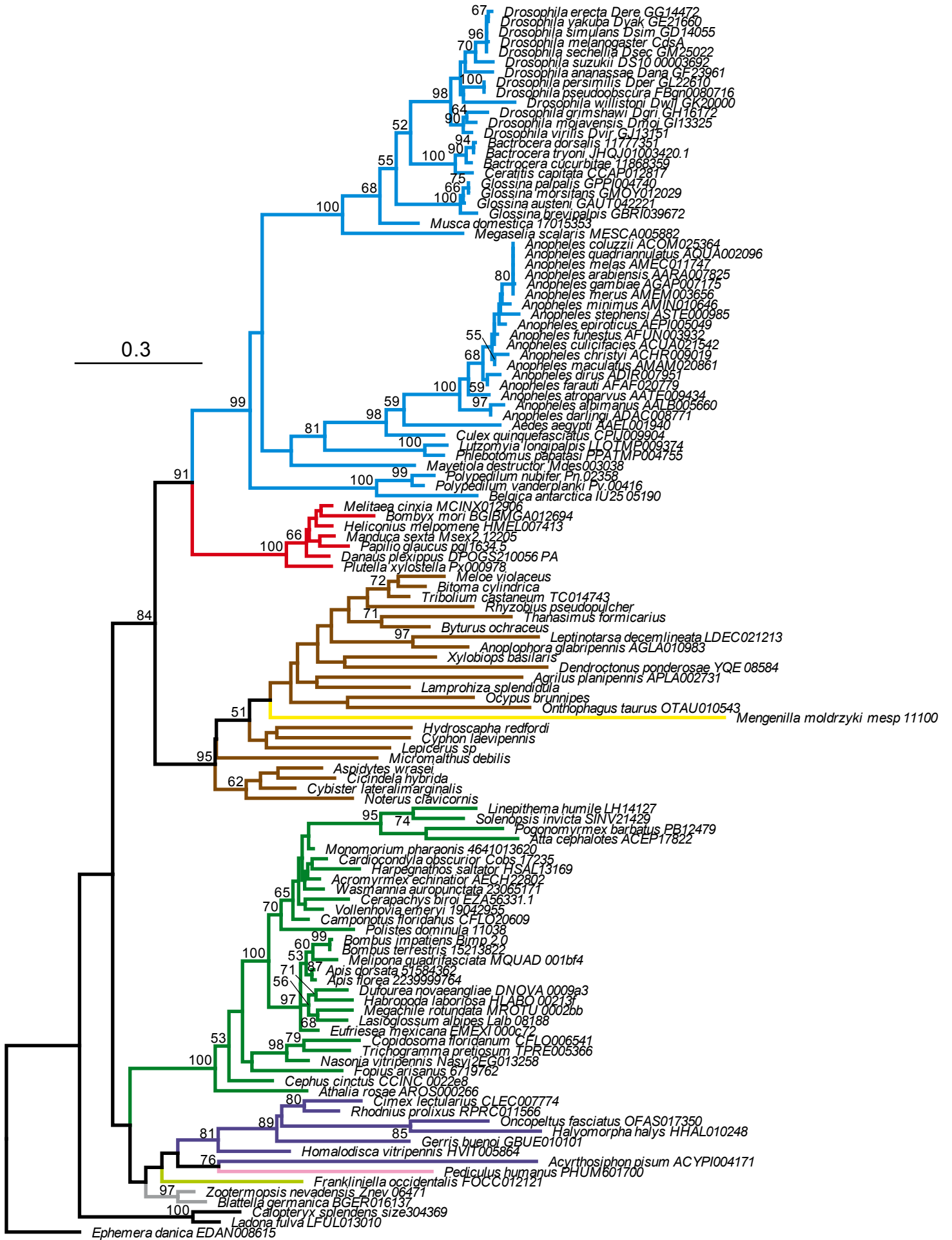




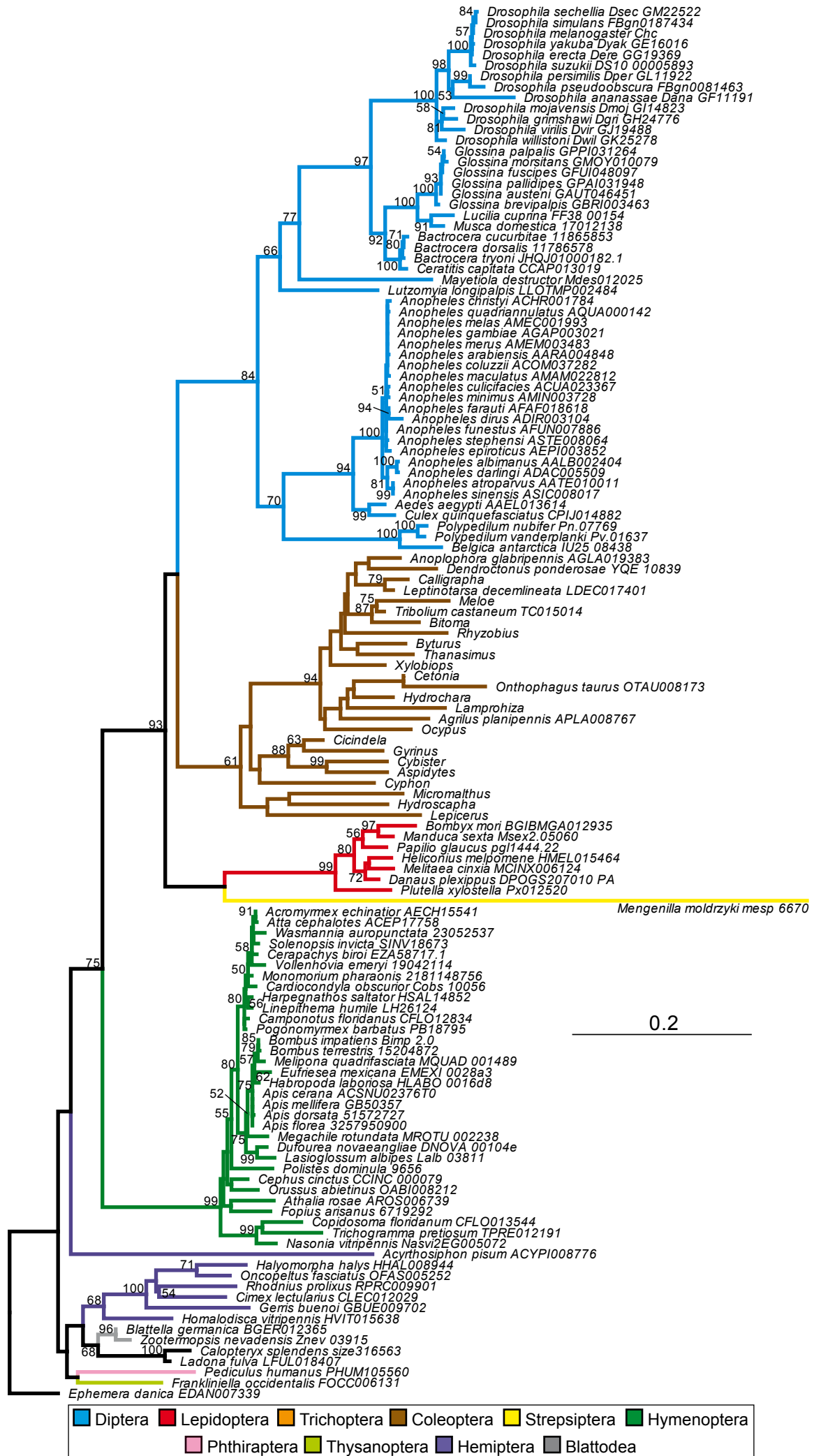
# Bug22



# CdsA

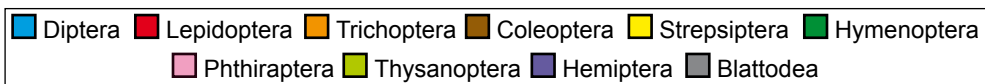
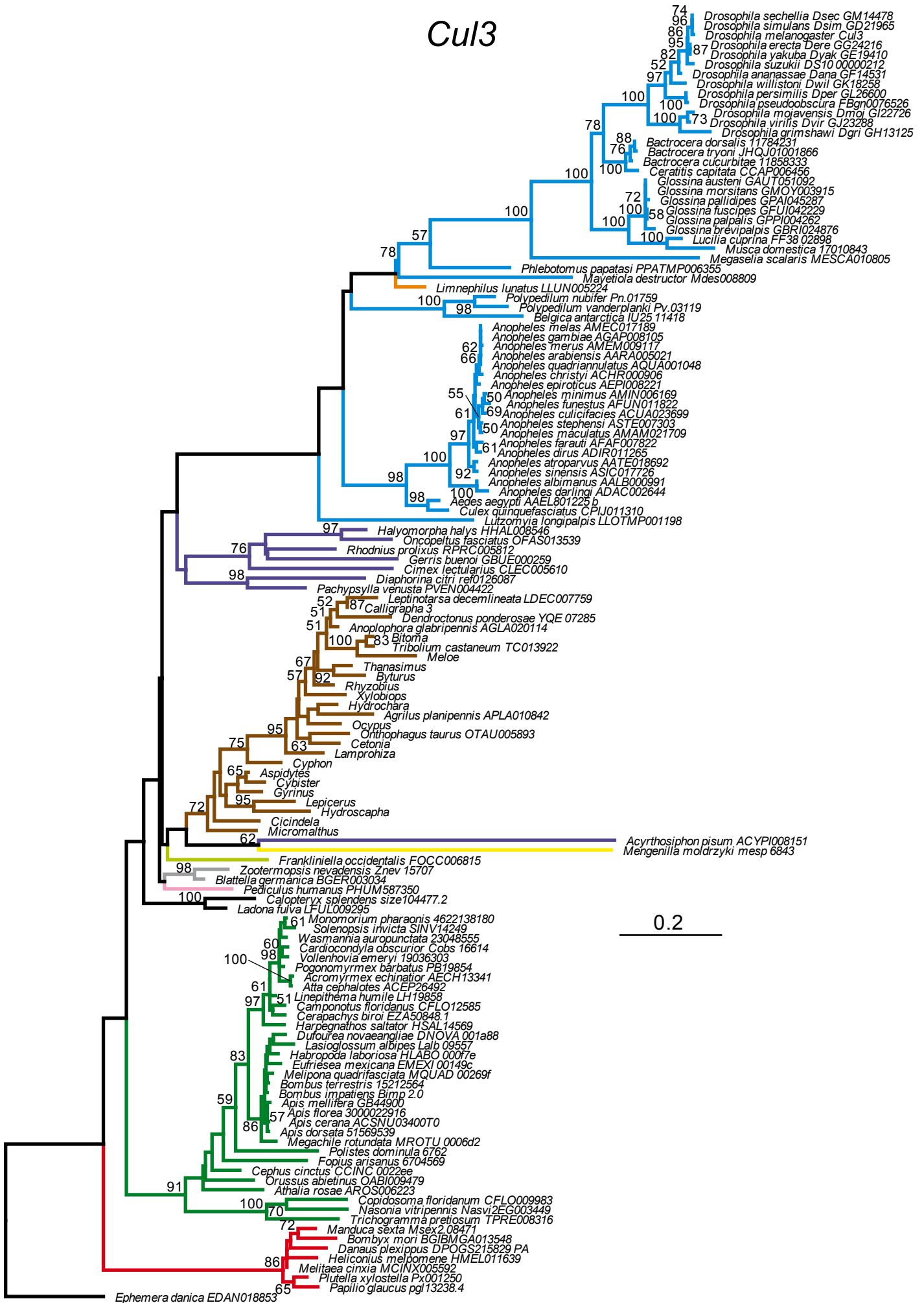


# Chc

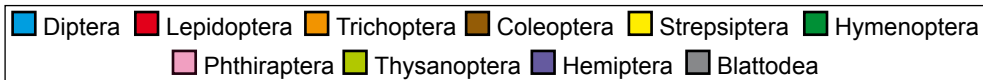
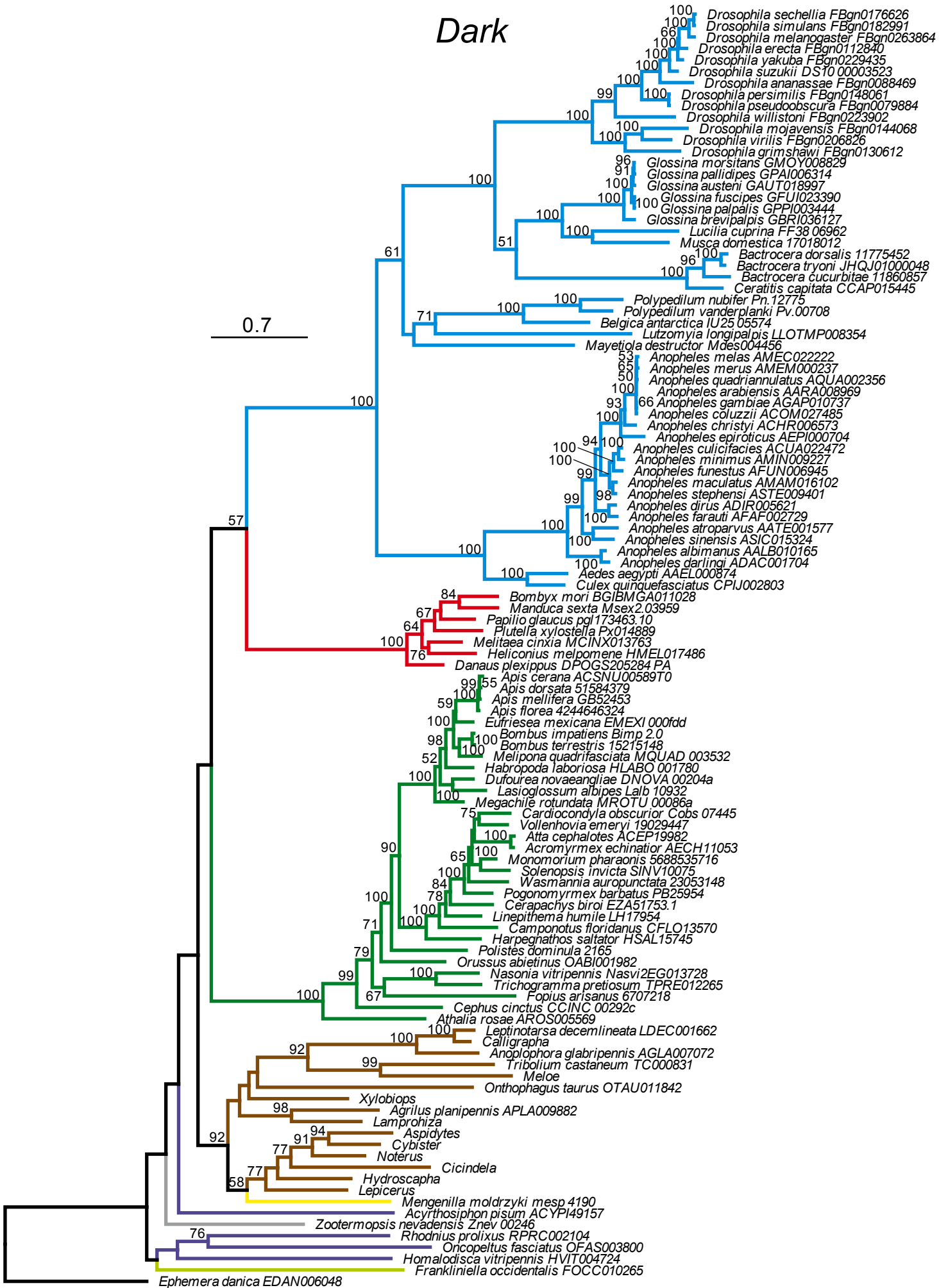




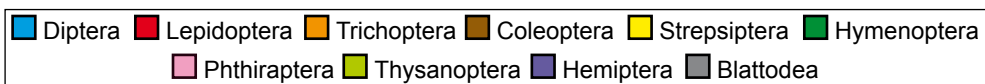
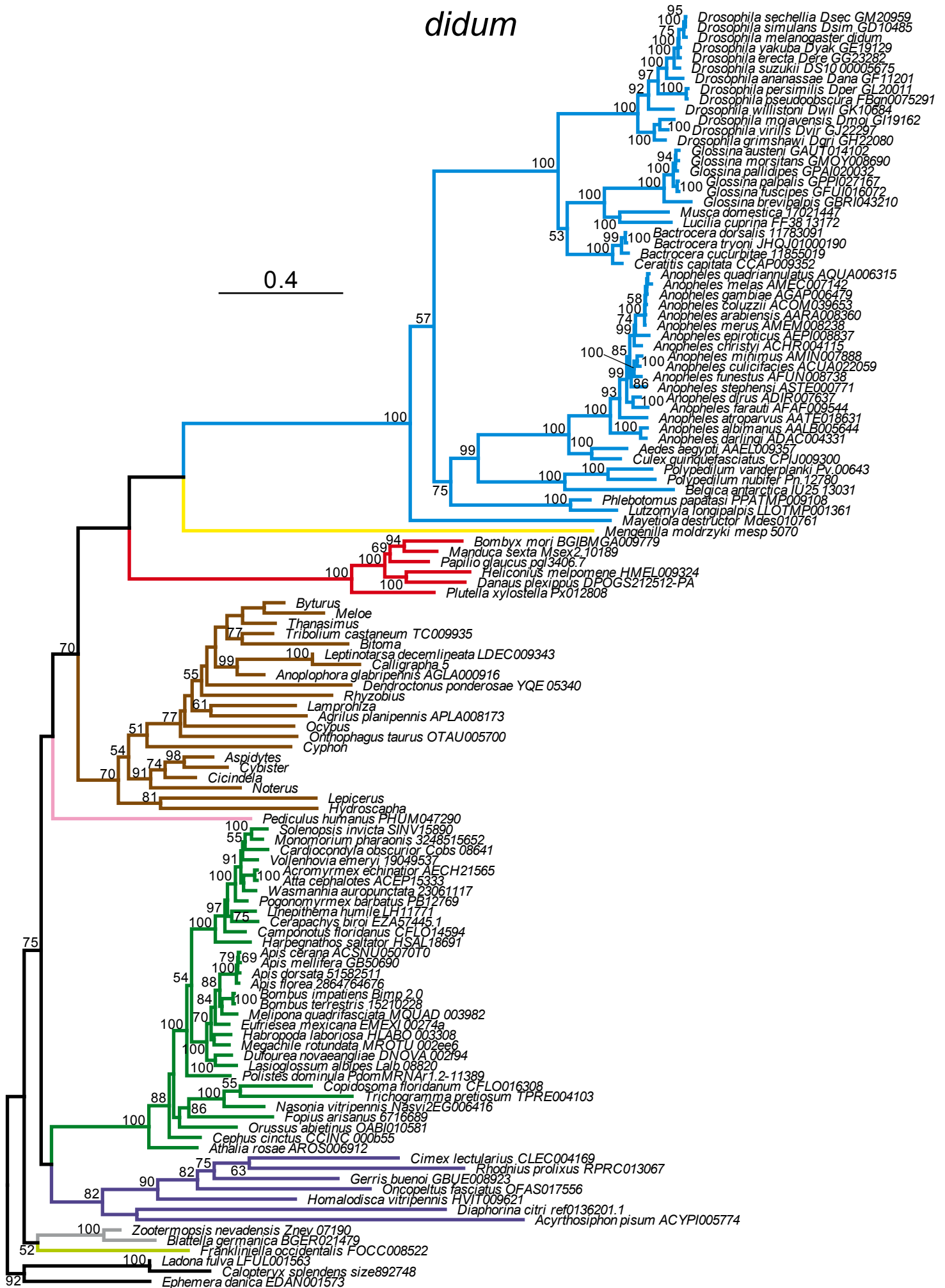
# Cul3



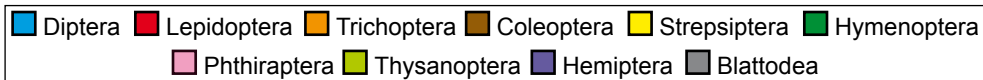
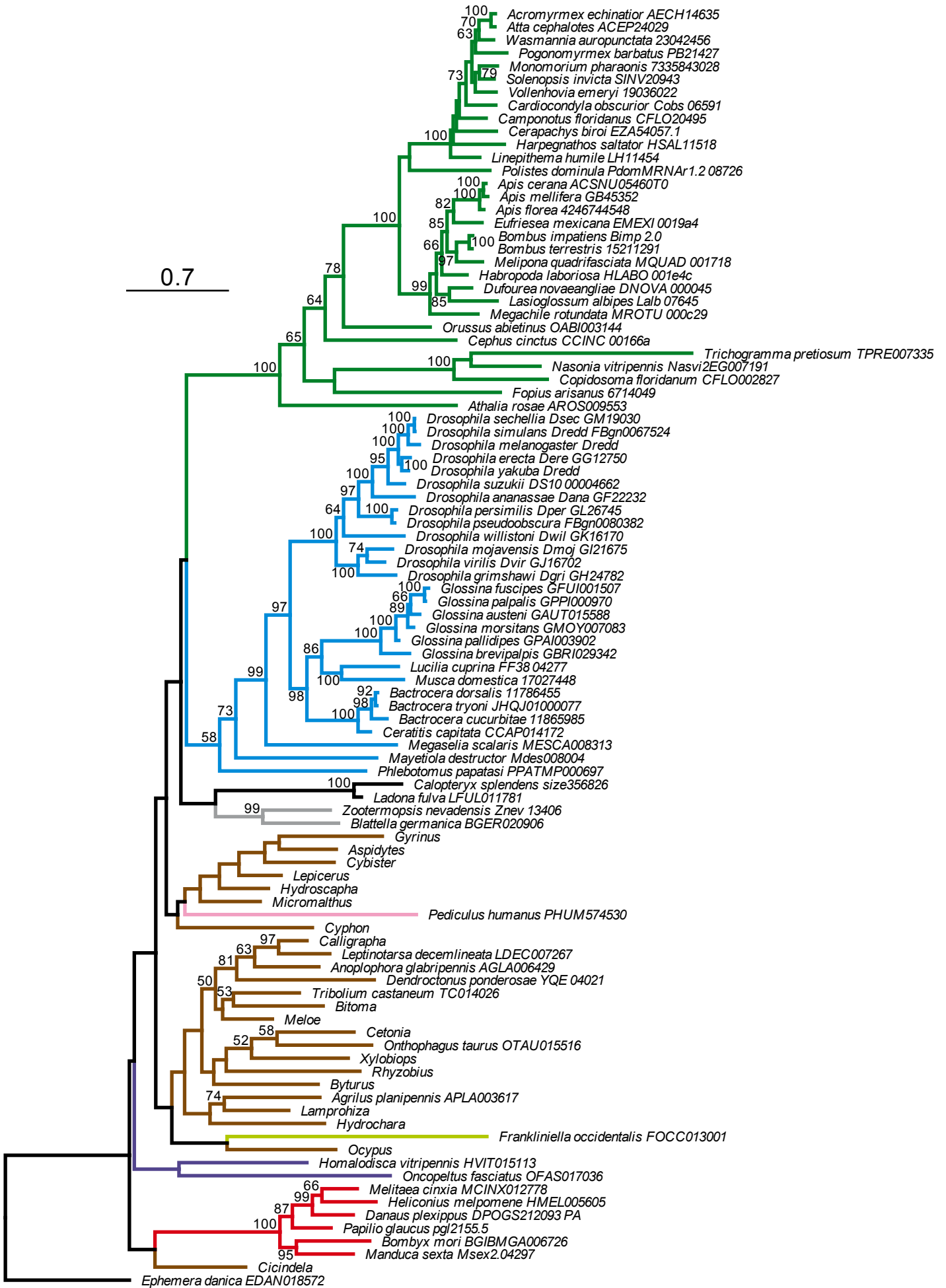
# Dark



# didum

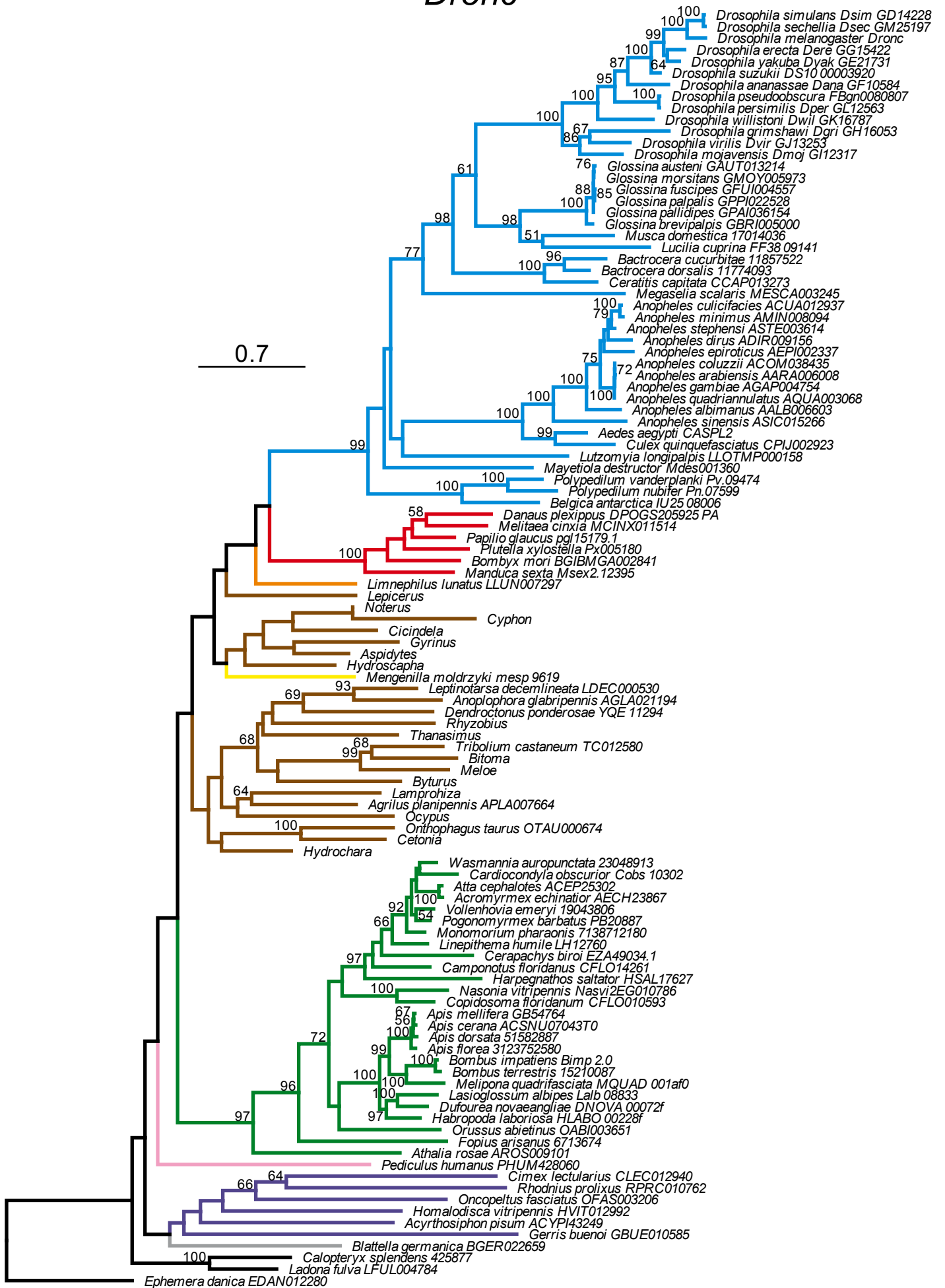


# Dredd

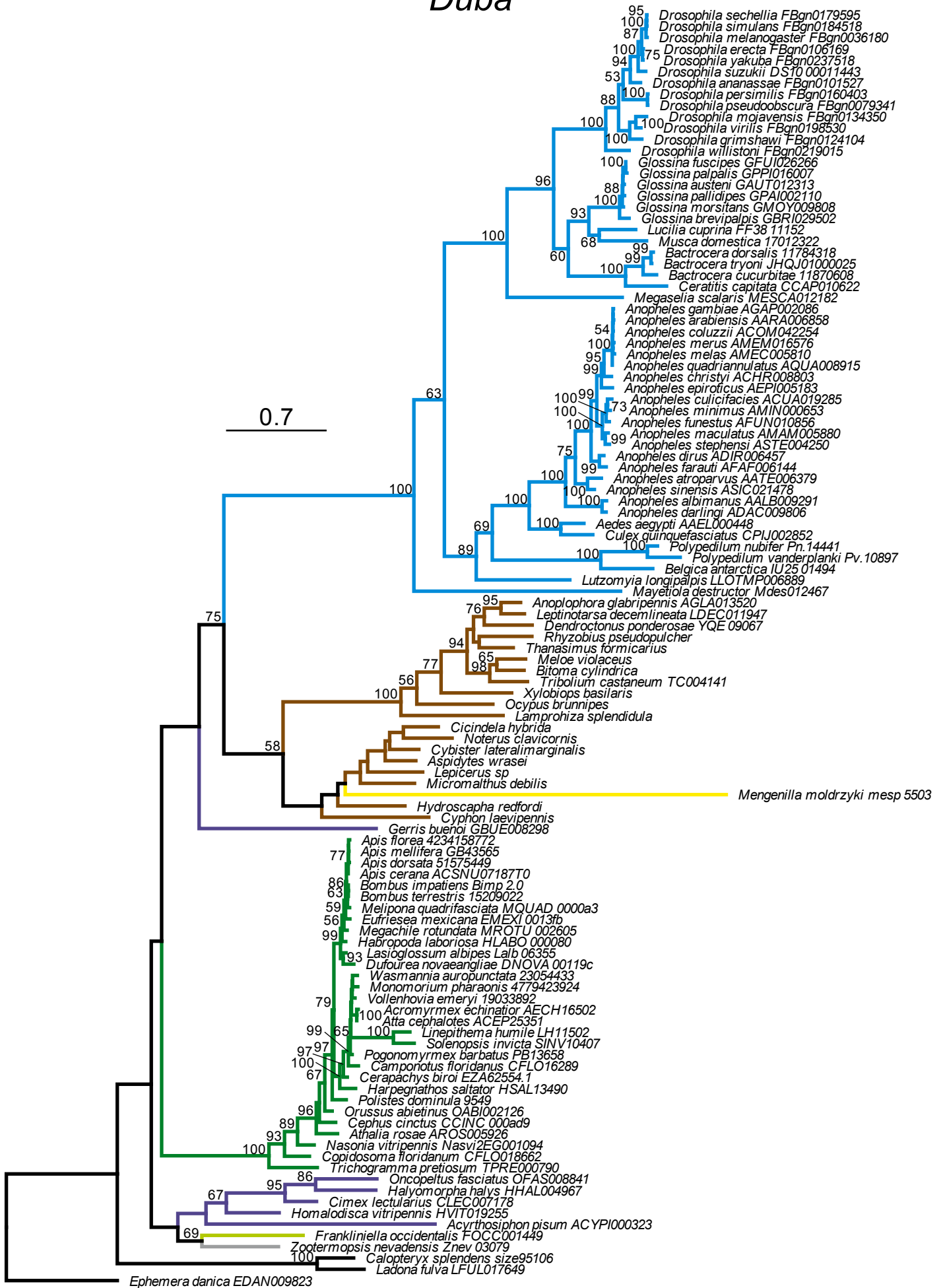




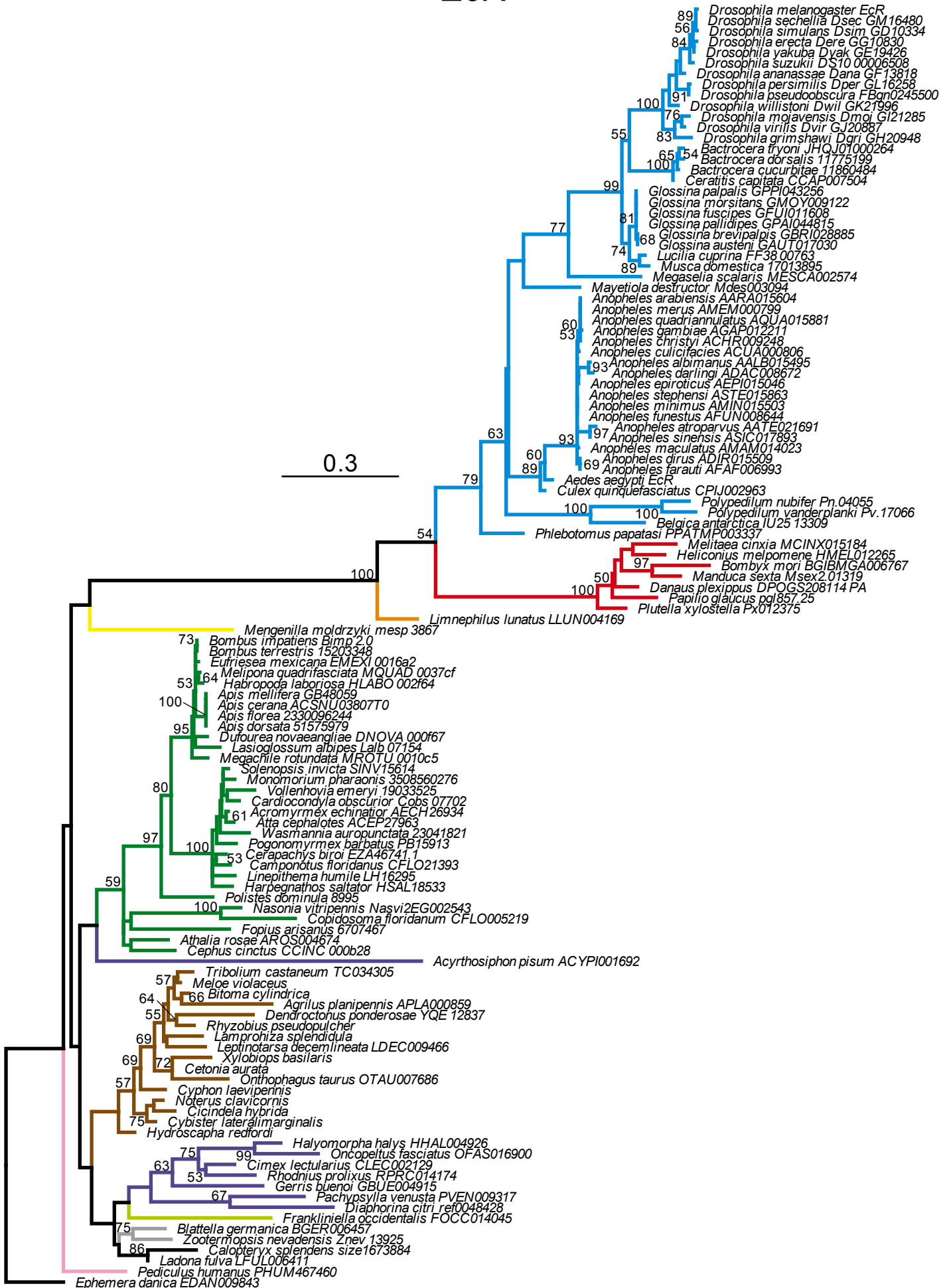
# Dronc



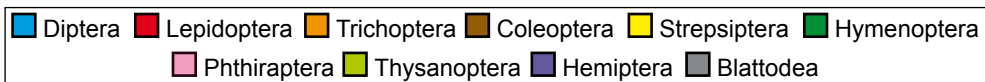
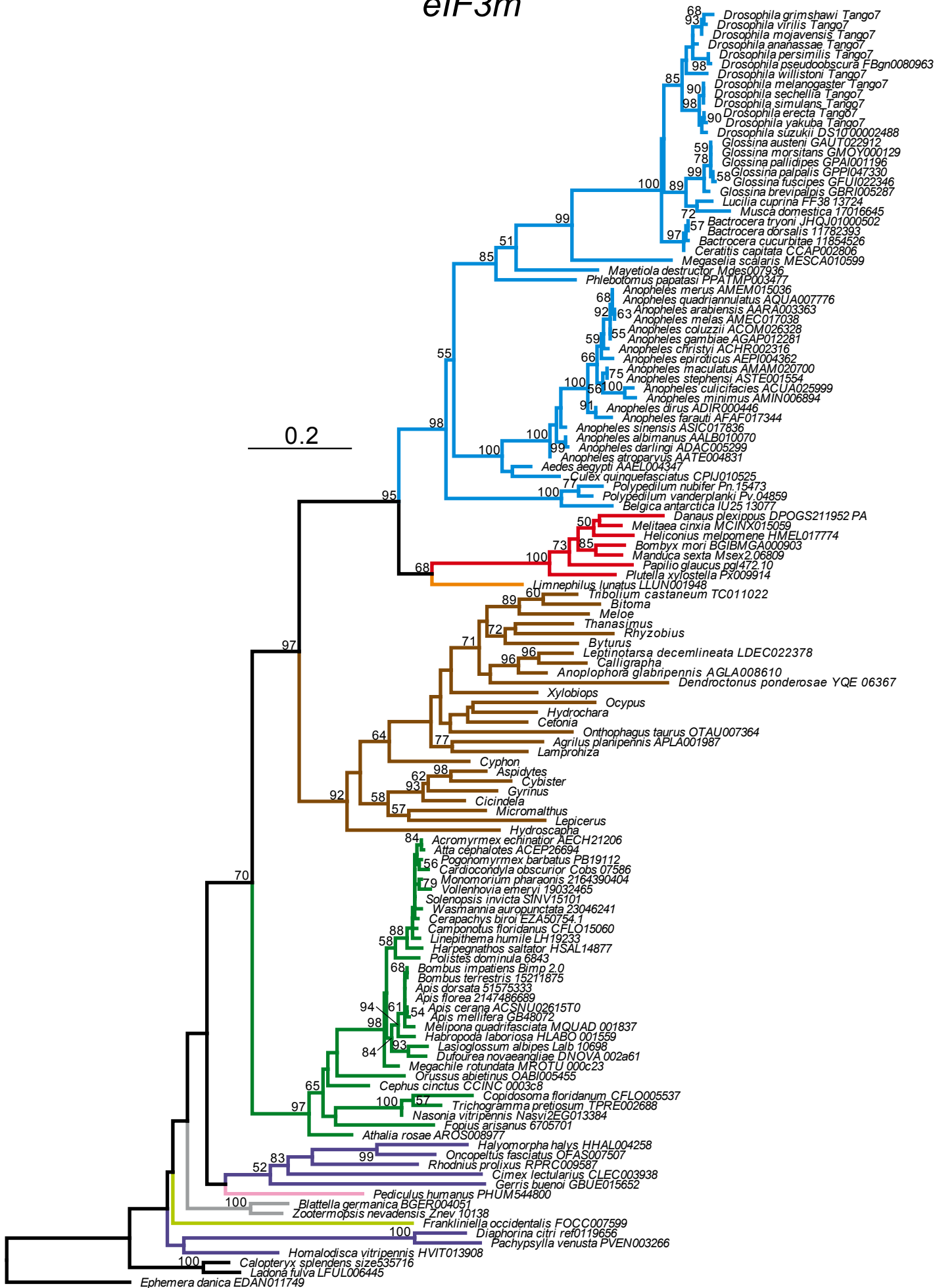
# Duba



# EcR

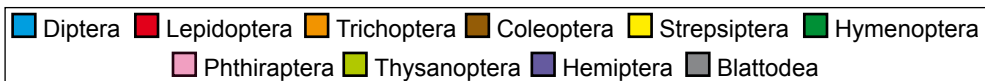
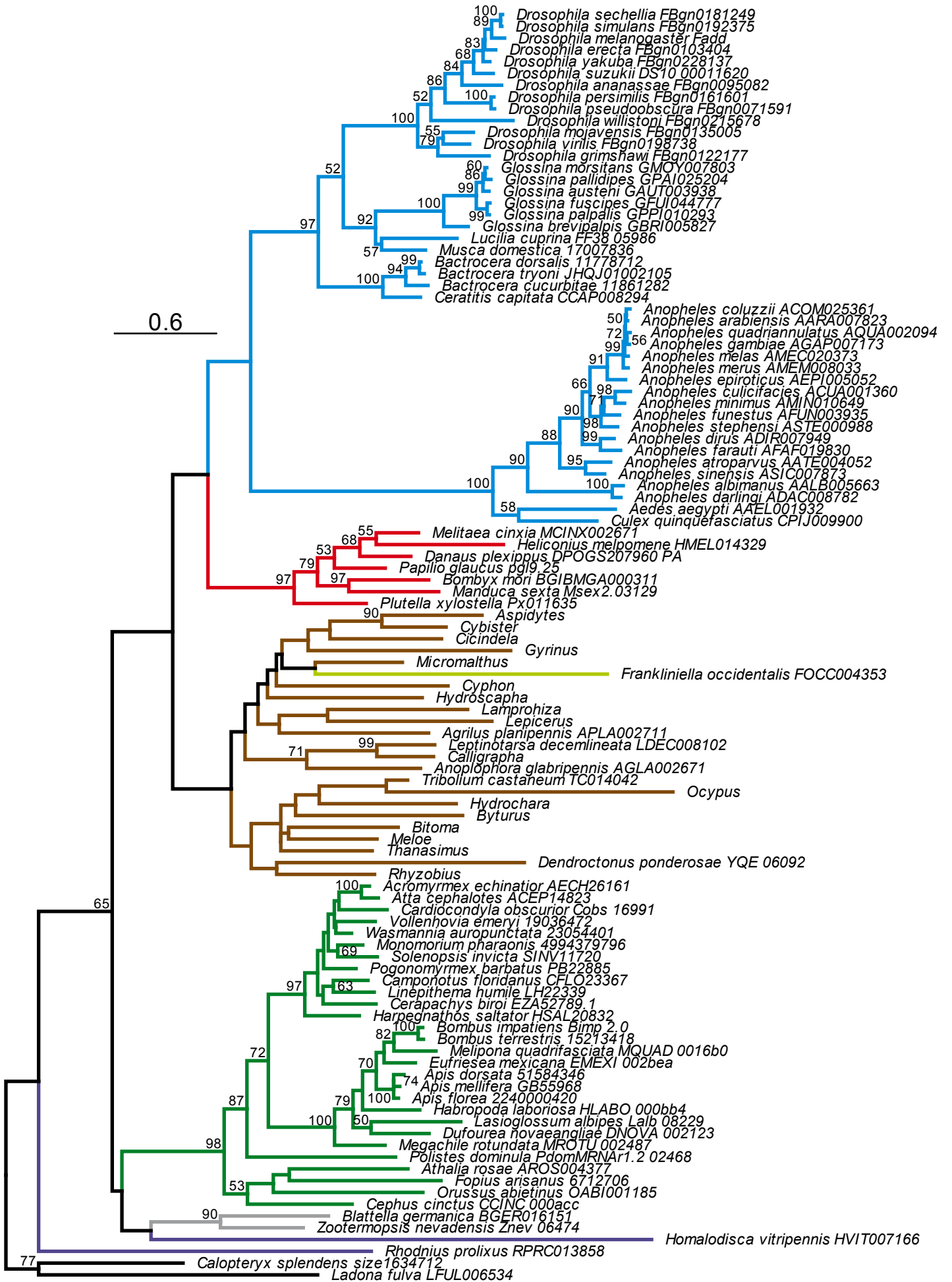


# eIF3m

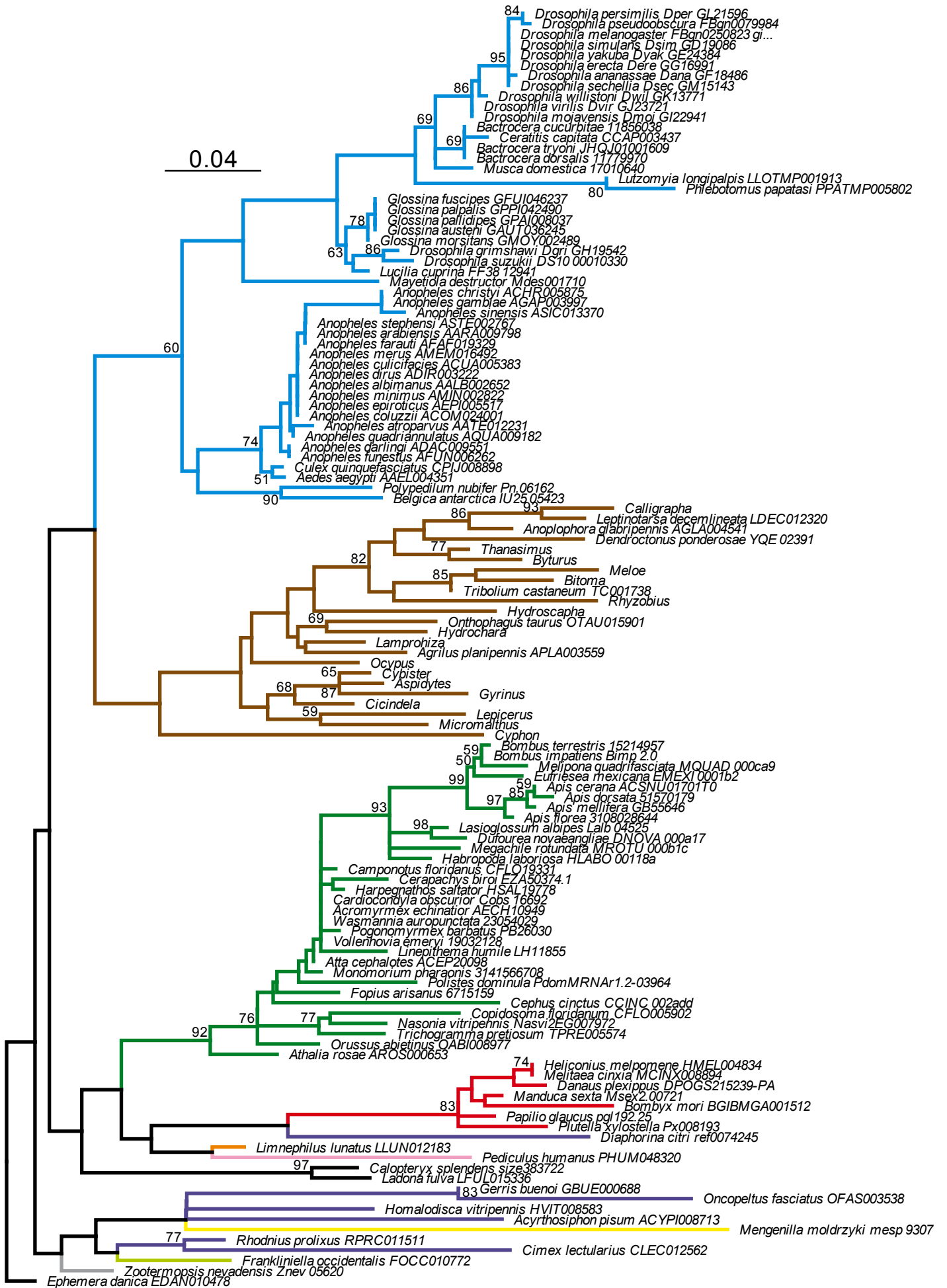




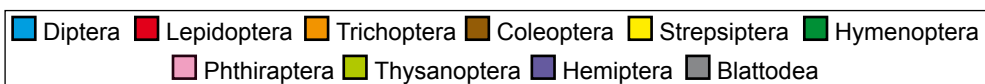
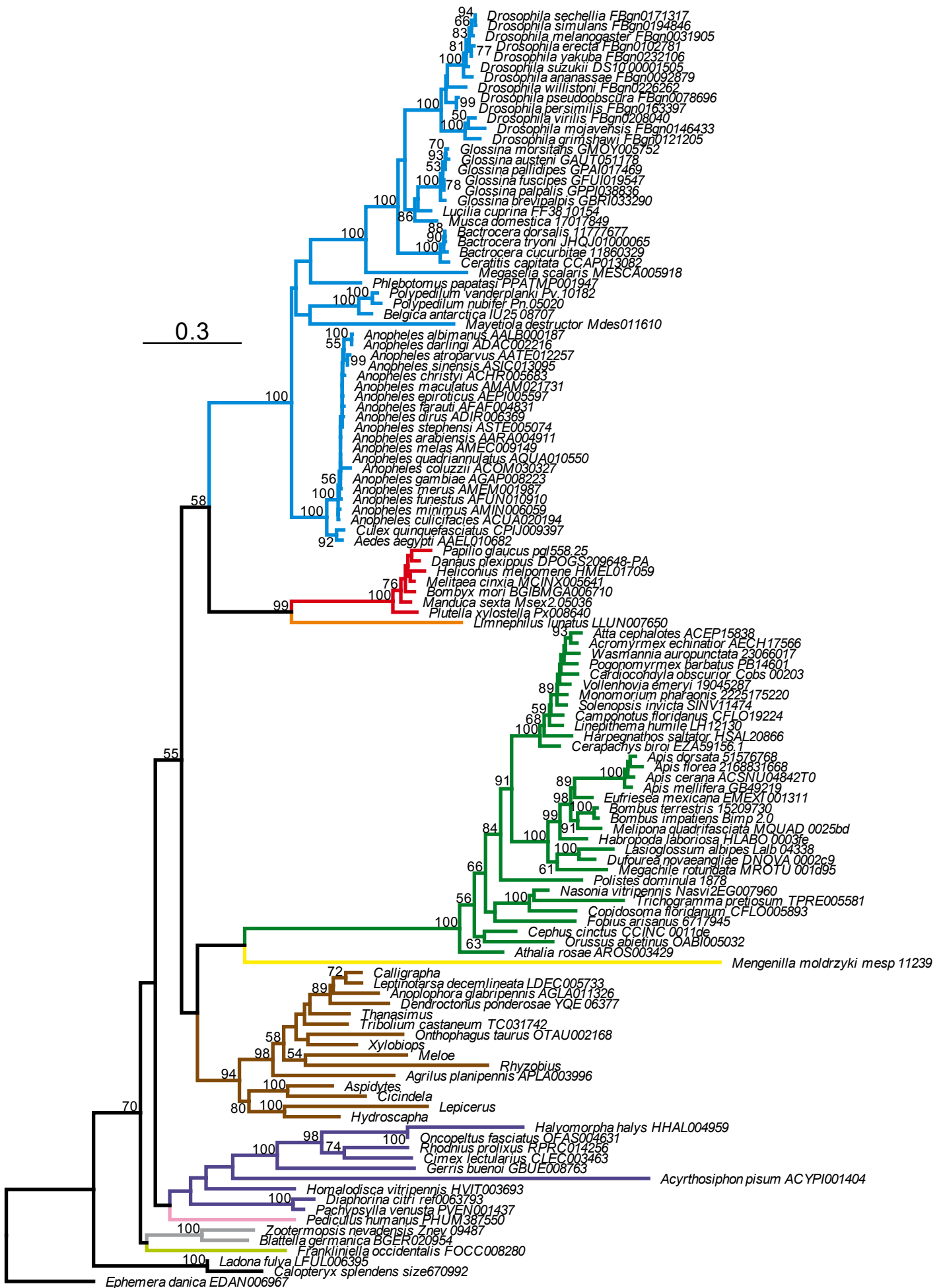
# Fadd



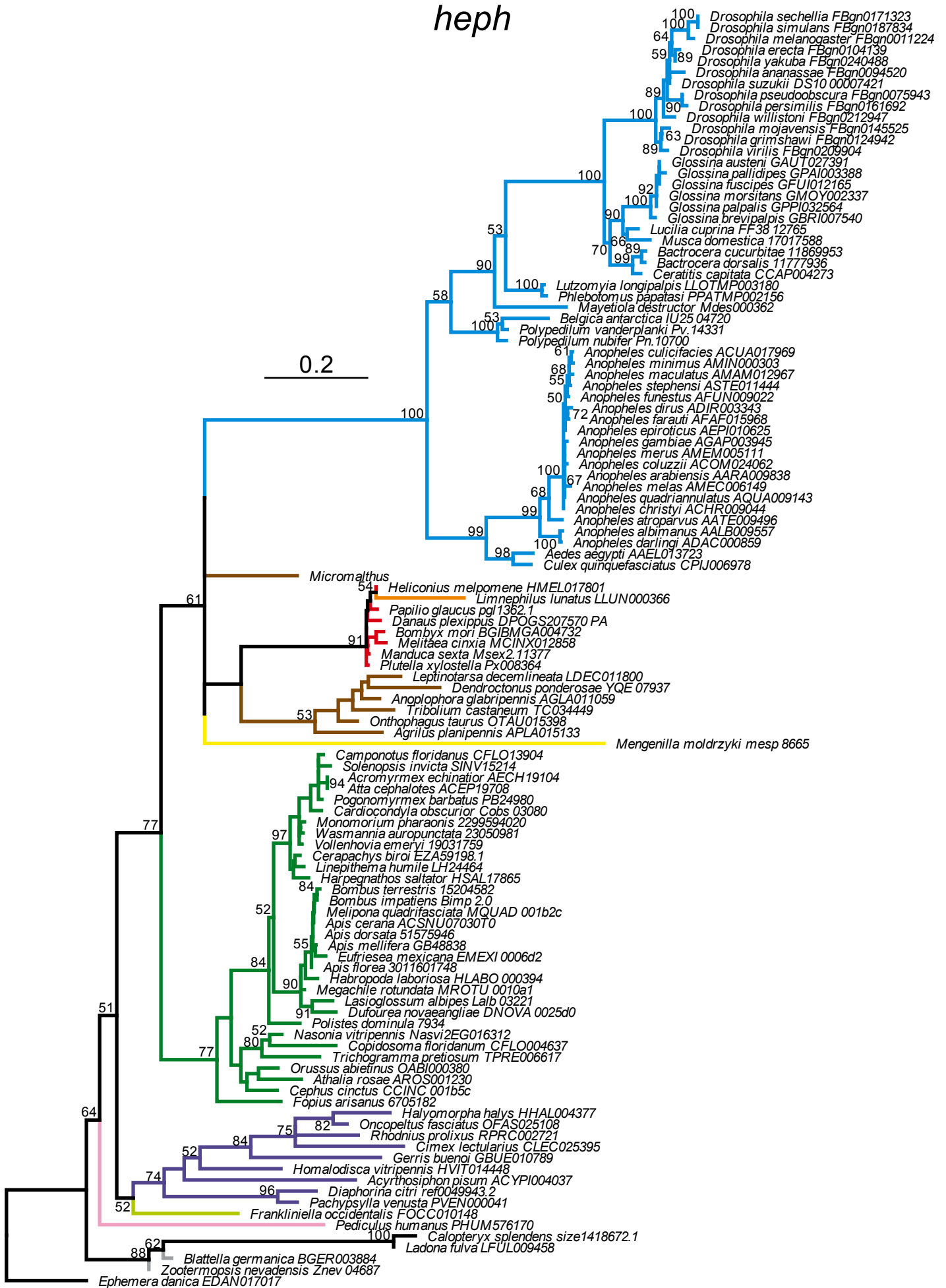
# gish



# gudu

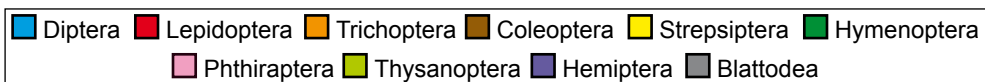
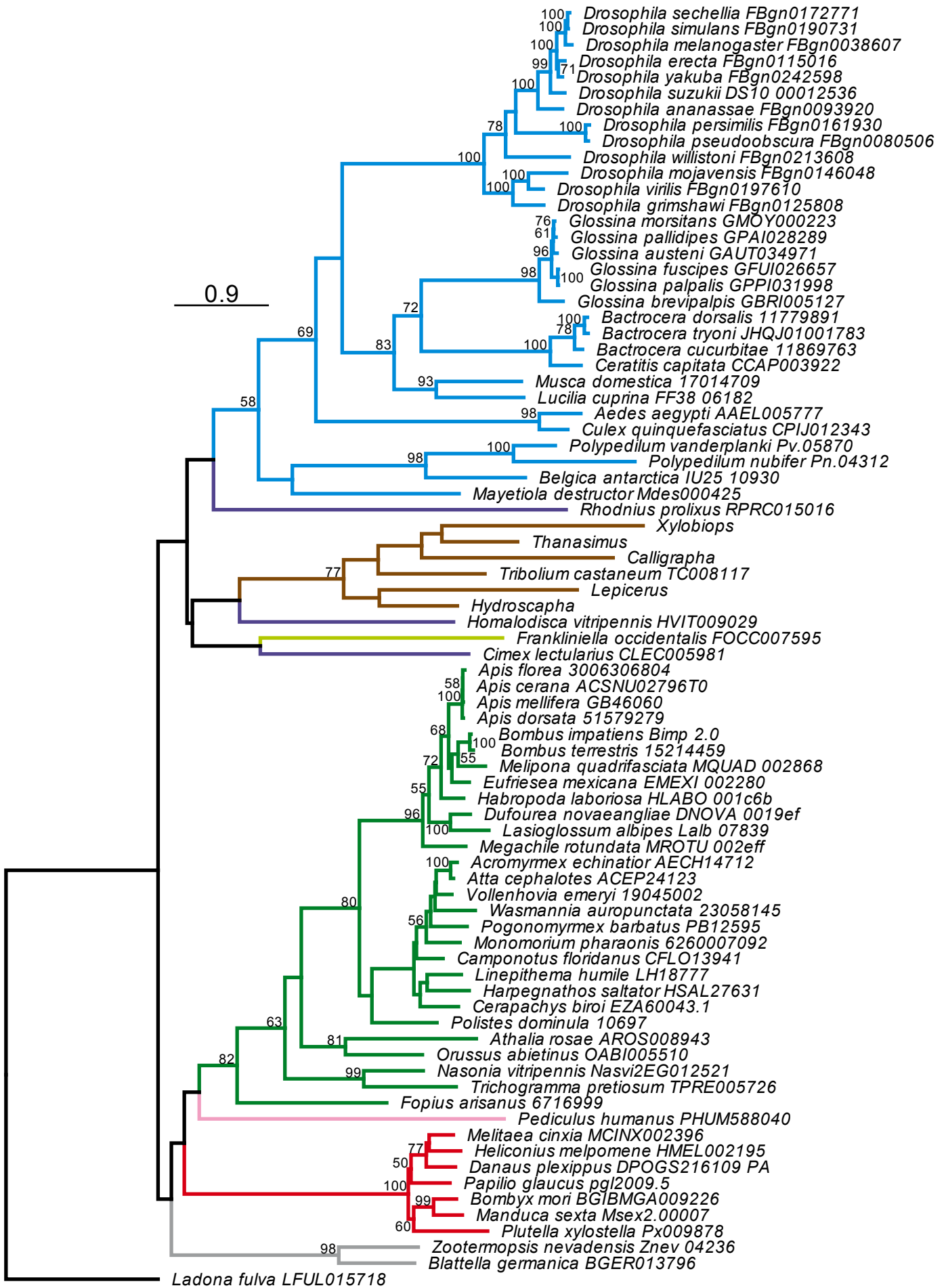


# heph

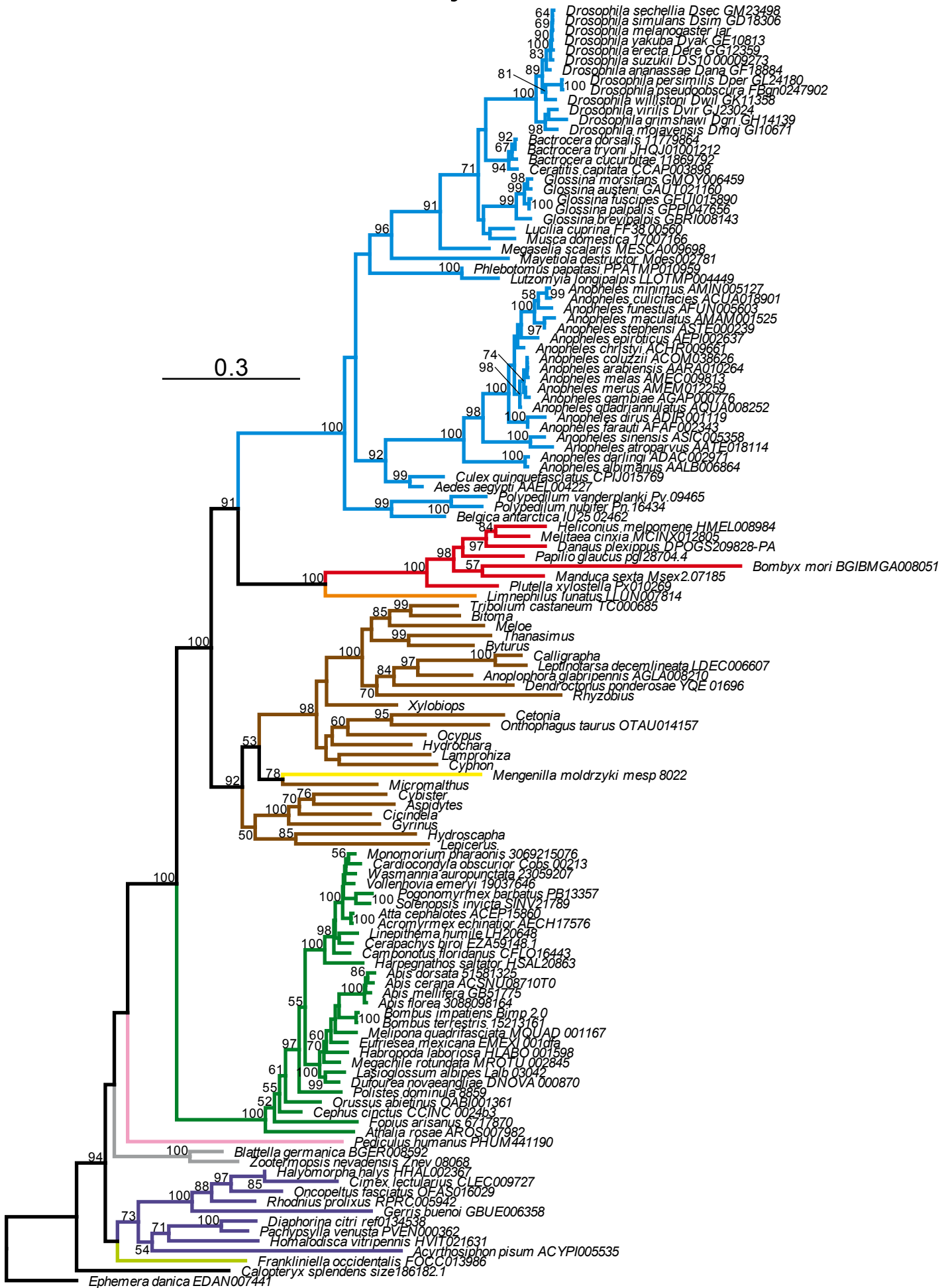




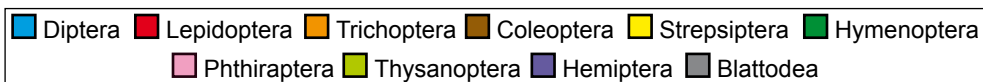
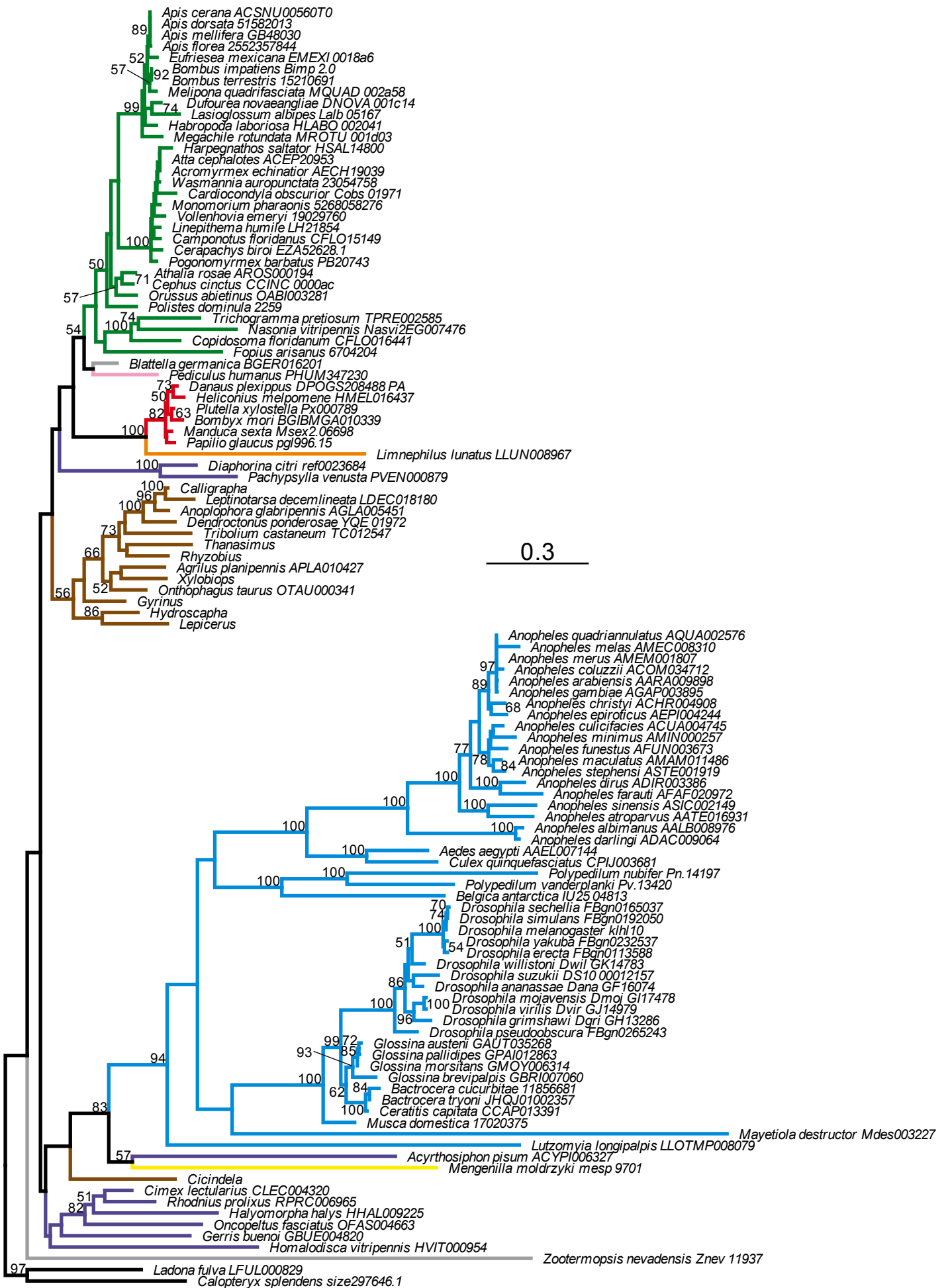
# hmw



jar

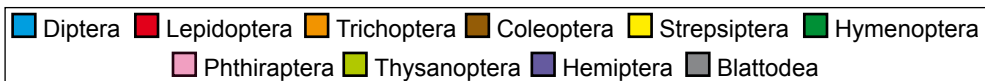
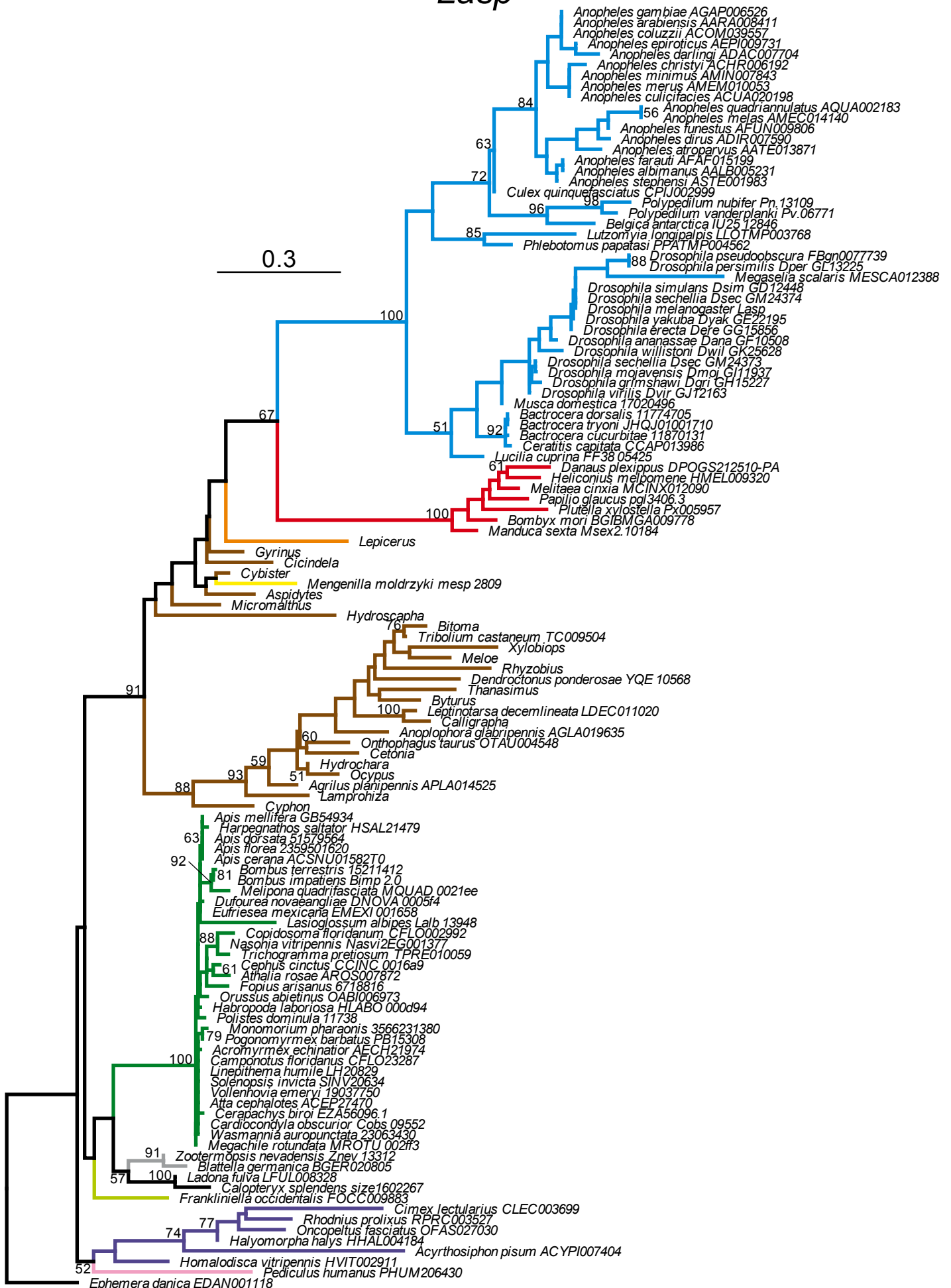


# klh10

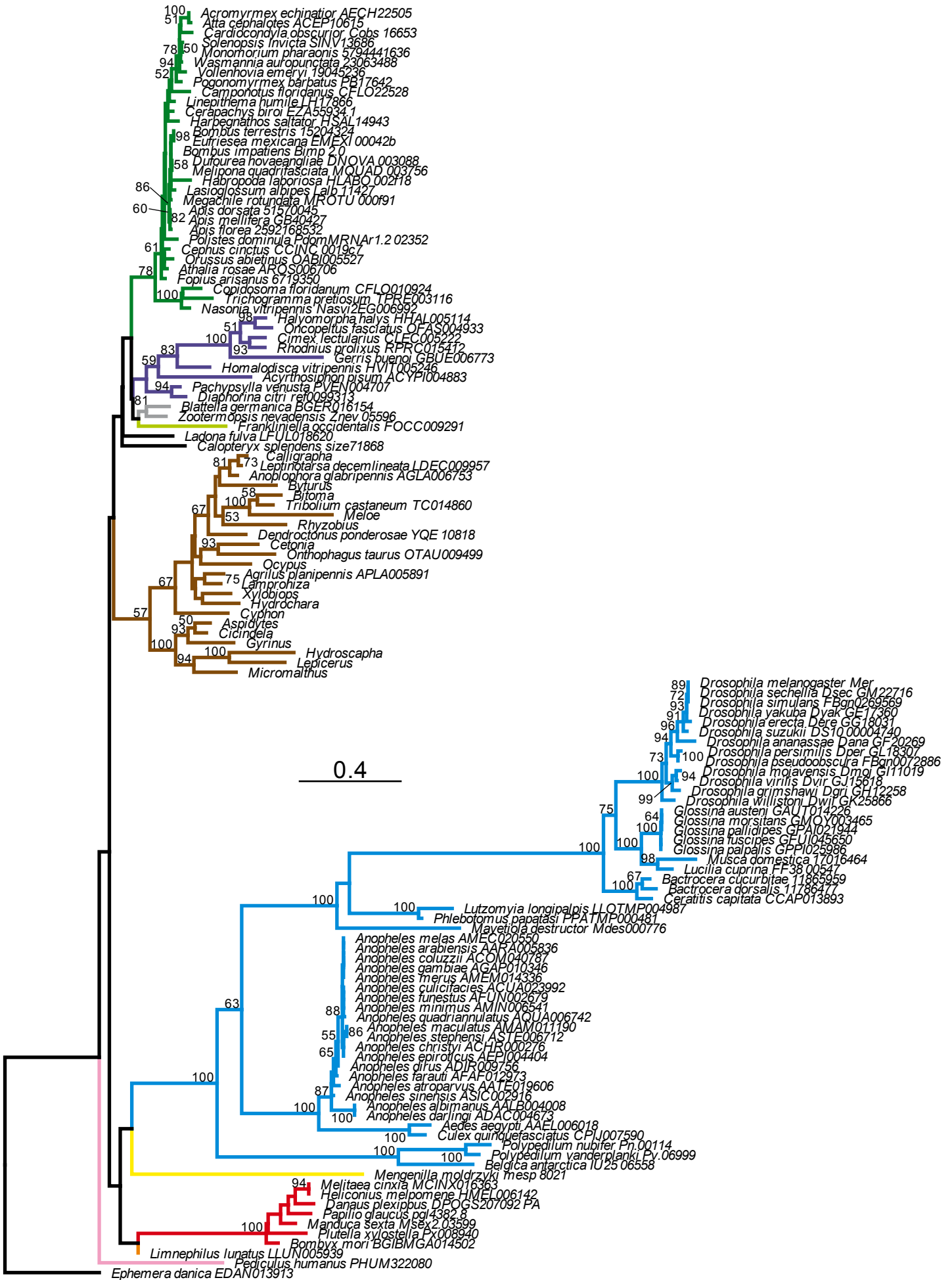




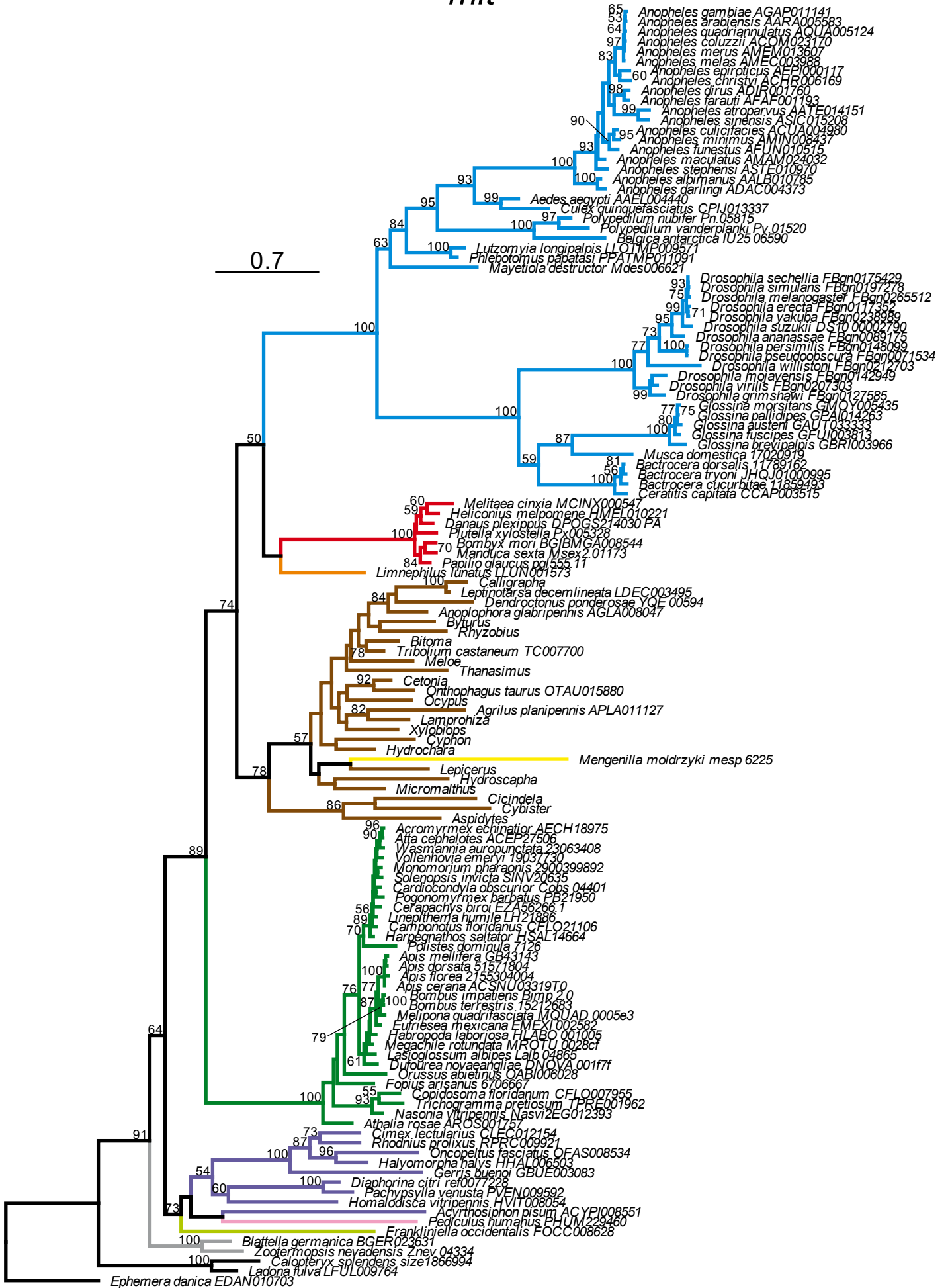
# Lasp



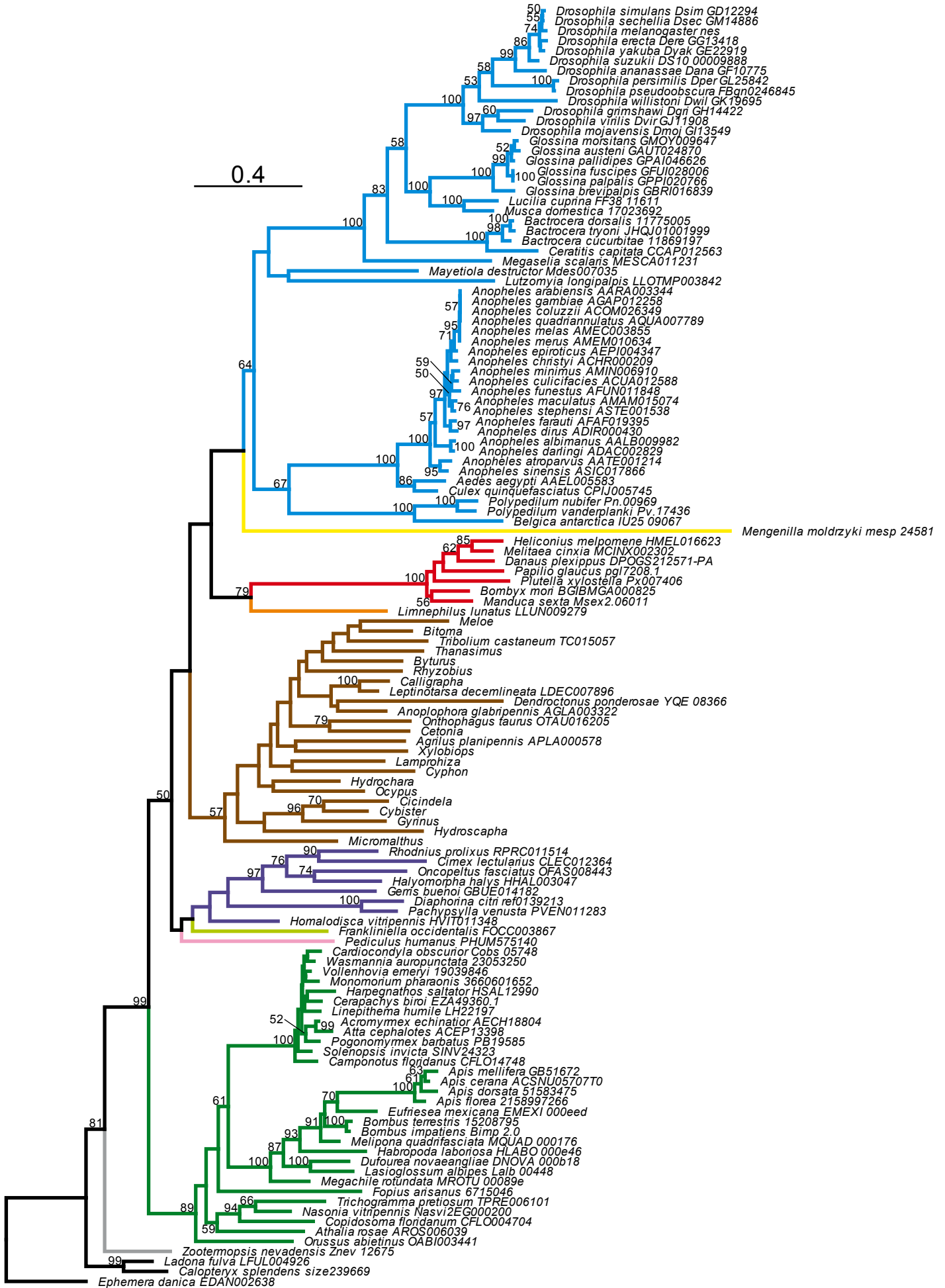
# Mer



# mit

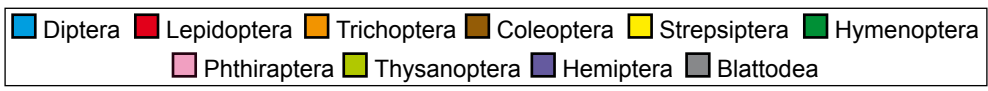
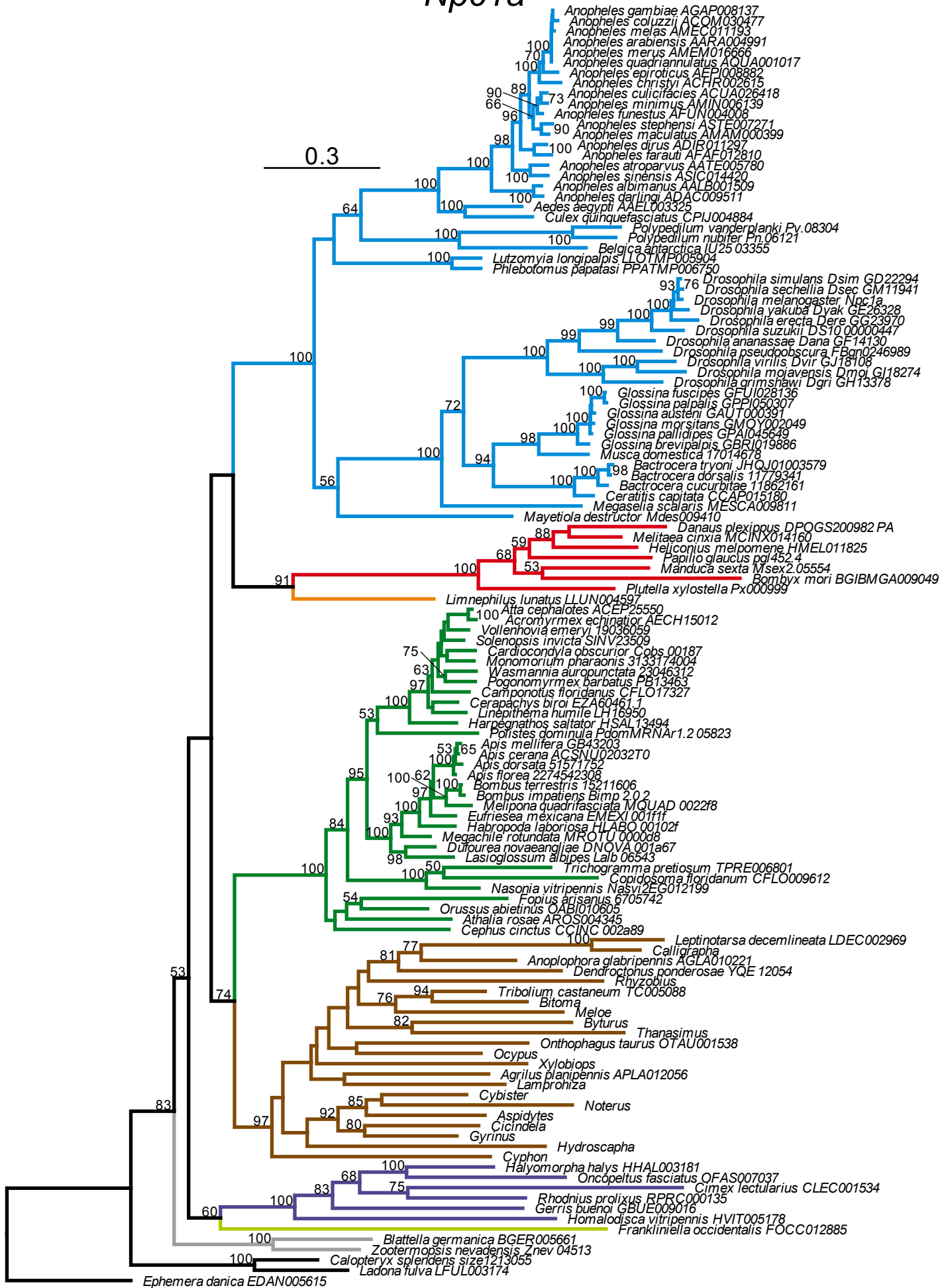


# nes

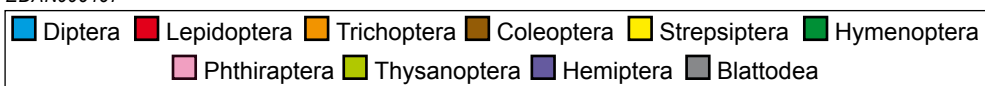
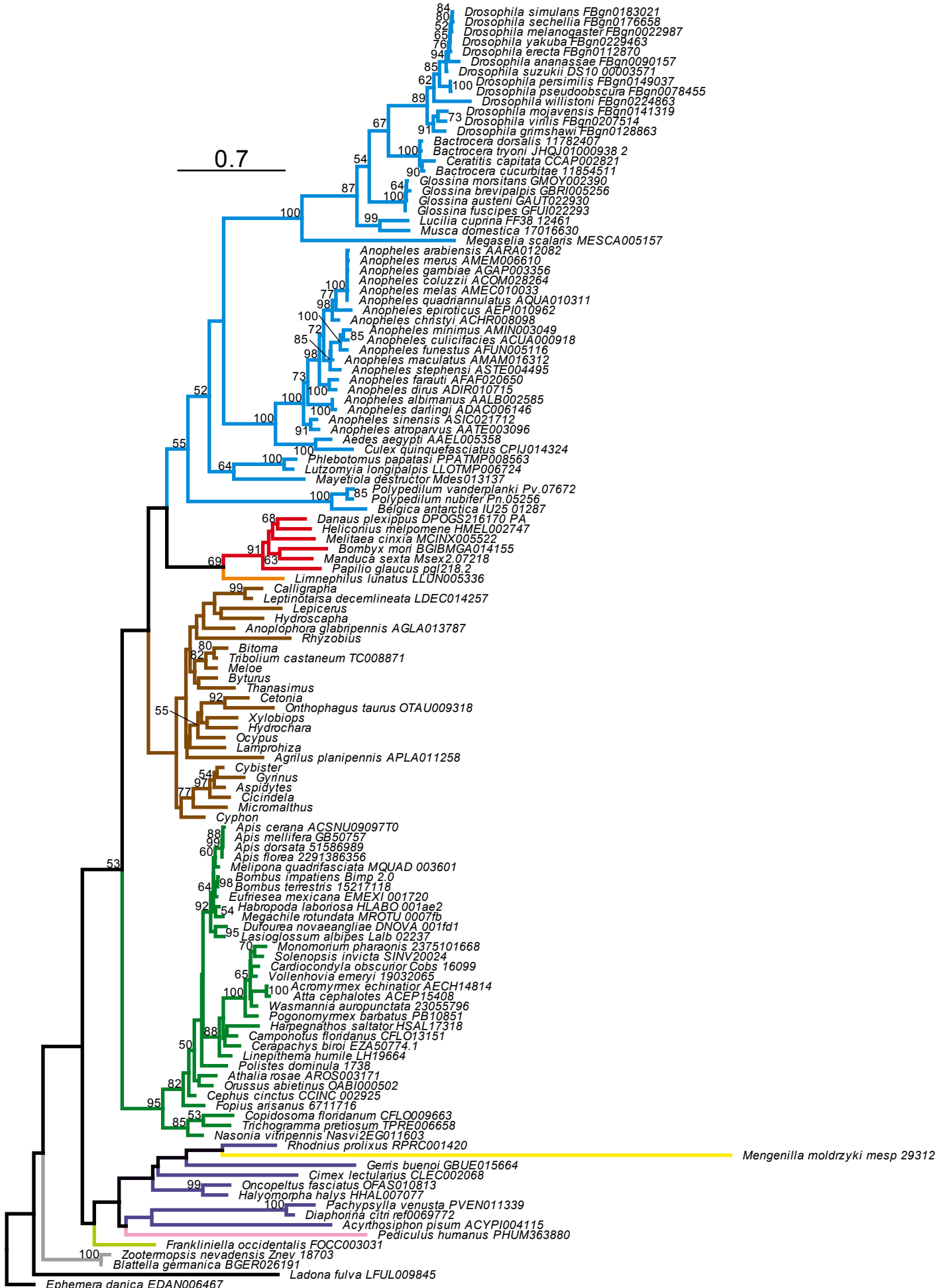




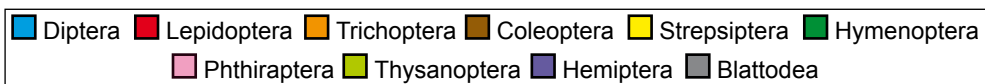
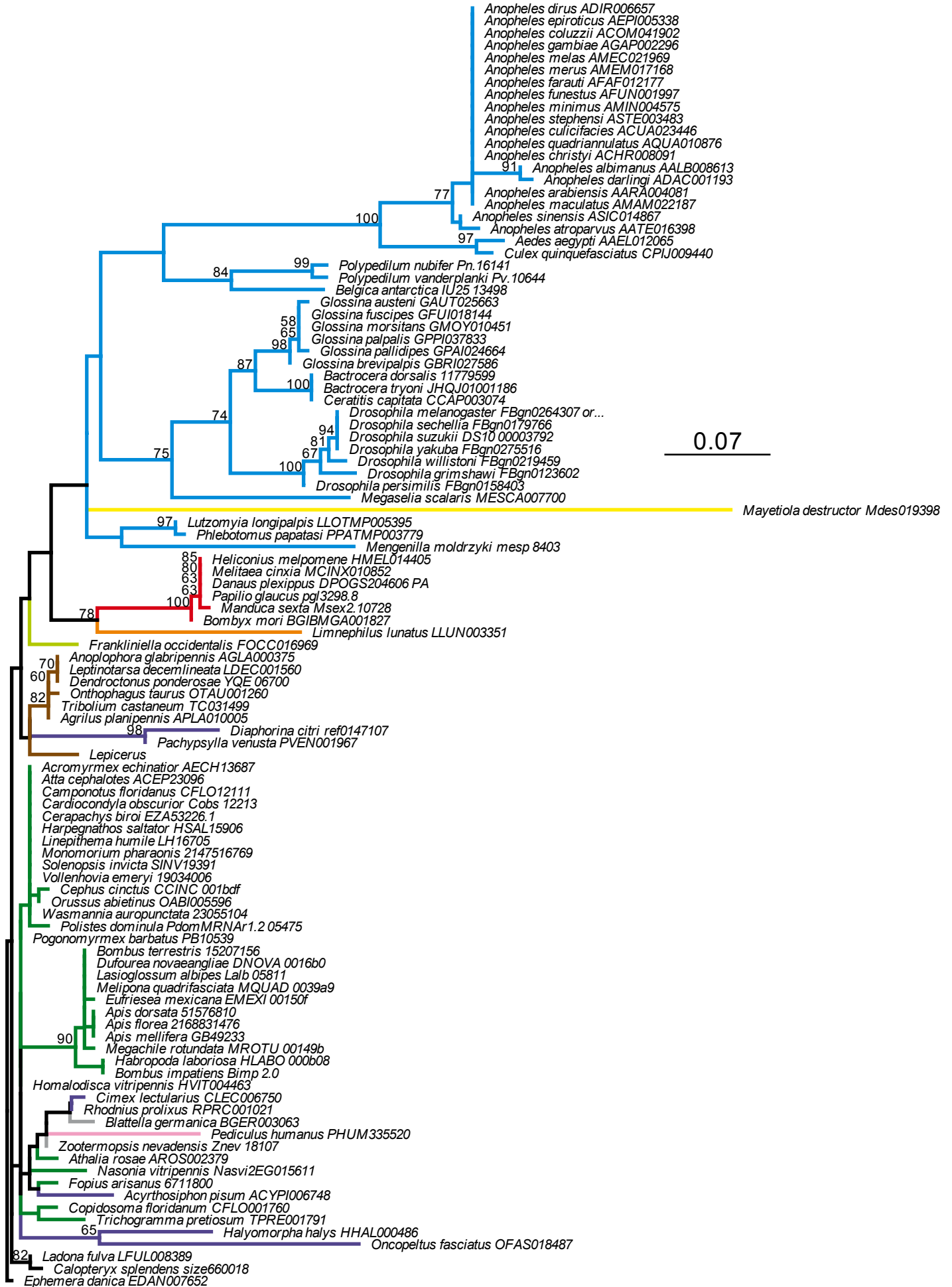
# Npc1a



nsr

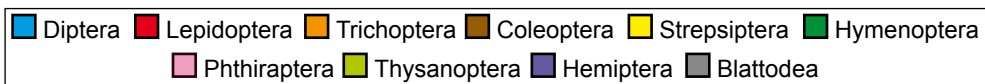
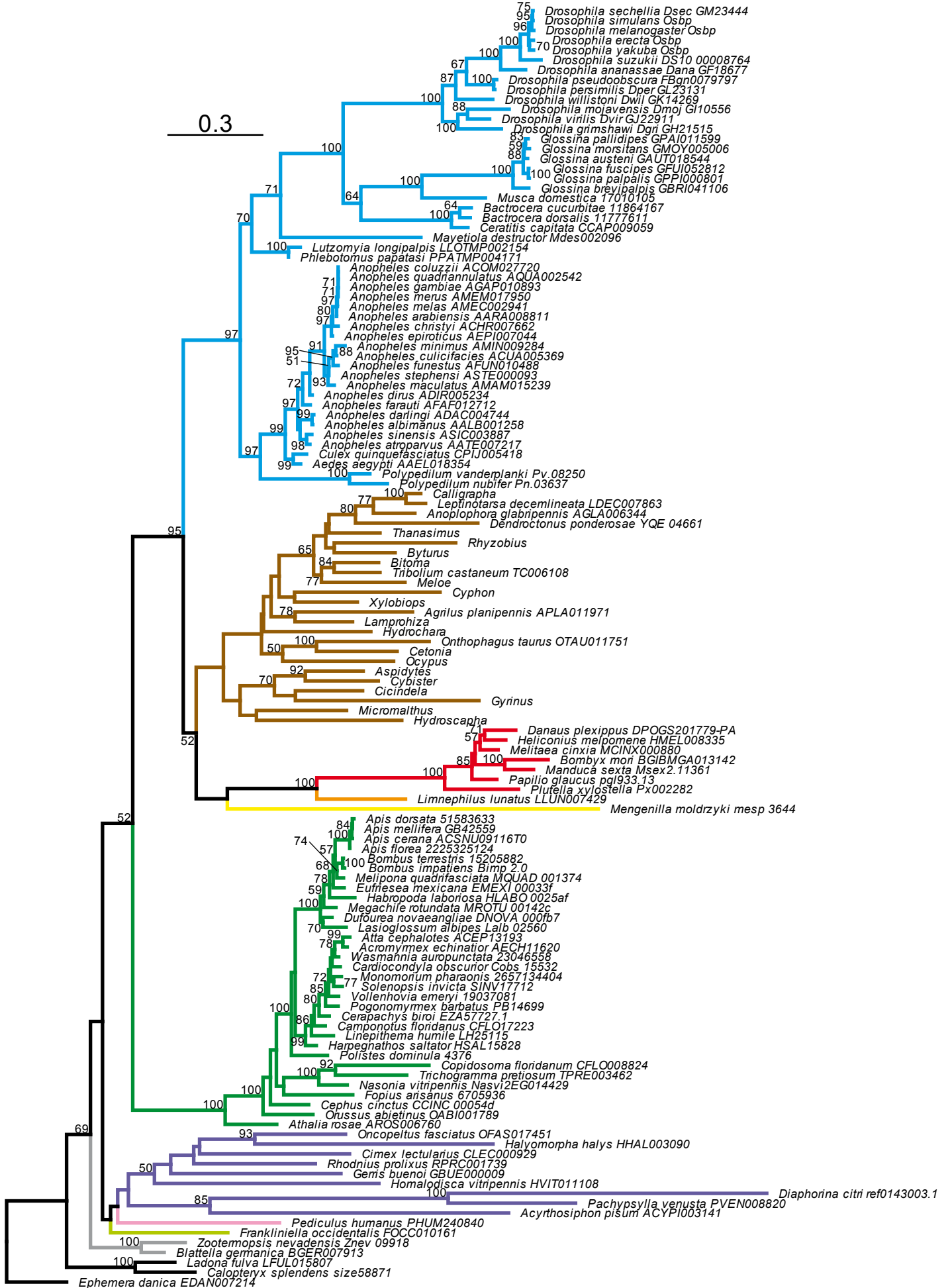


# orb2

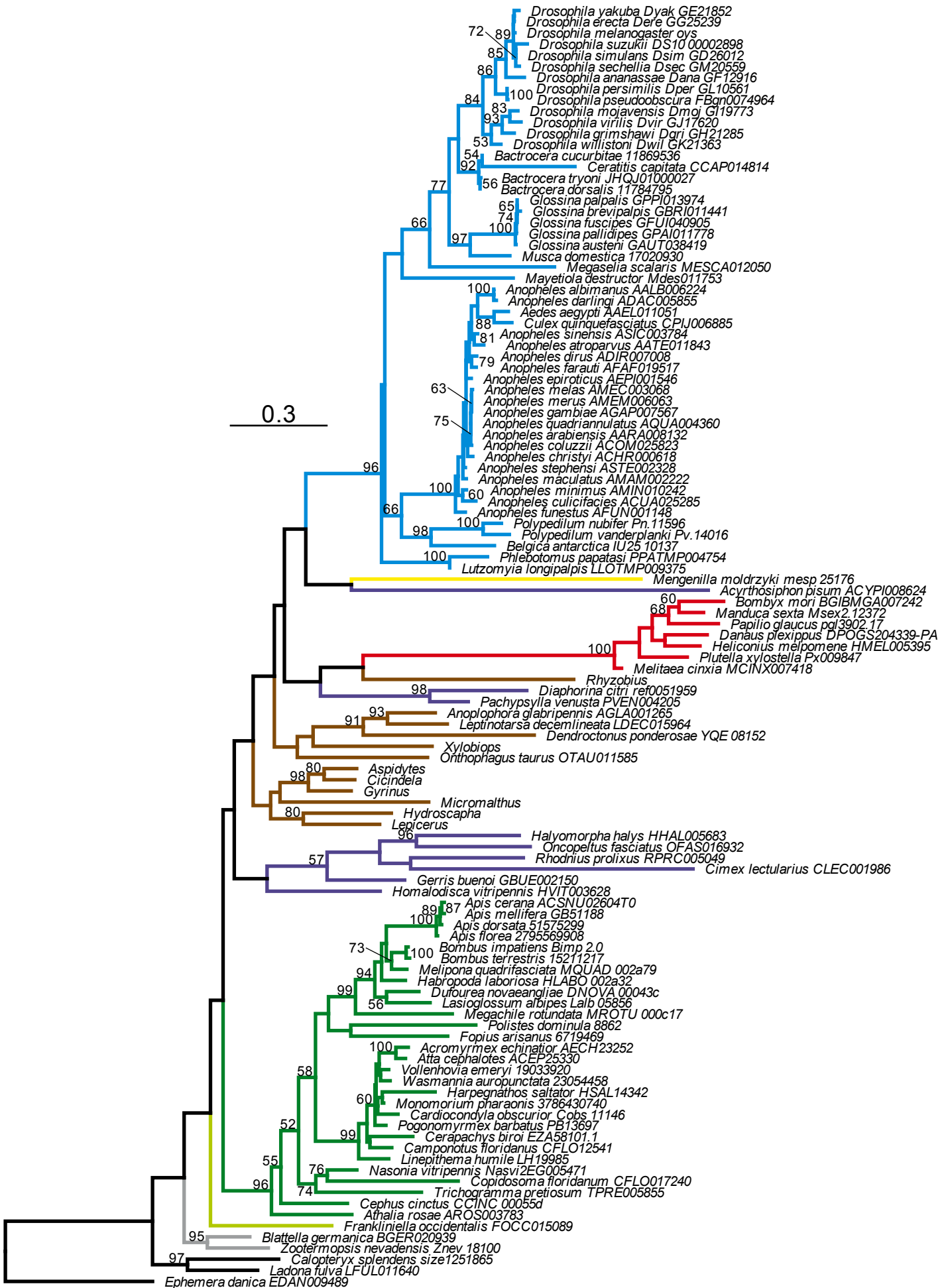




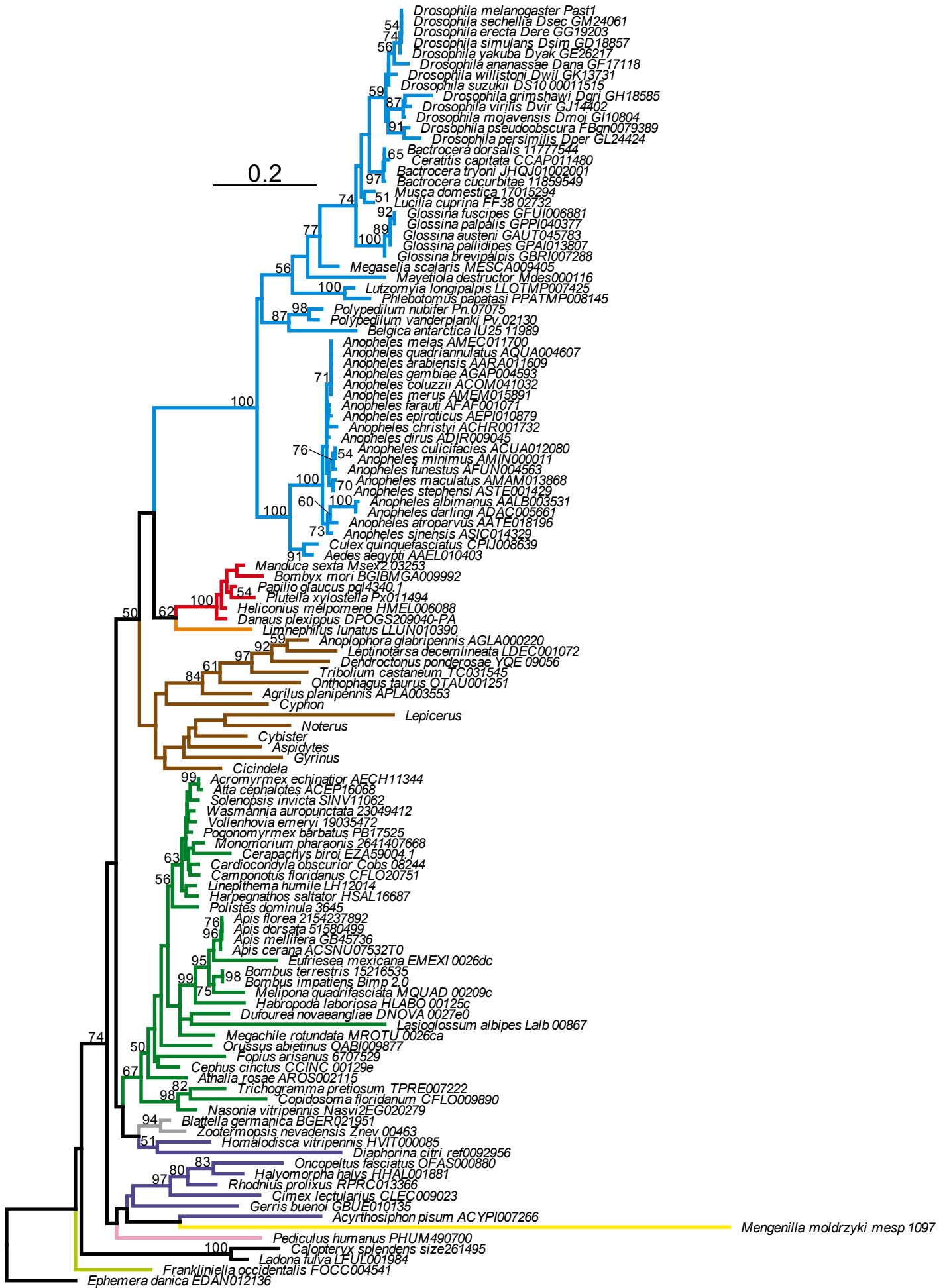
# Osbp



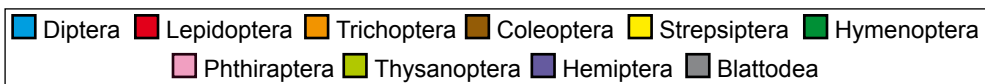
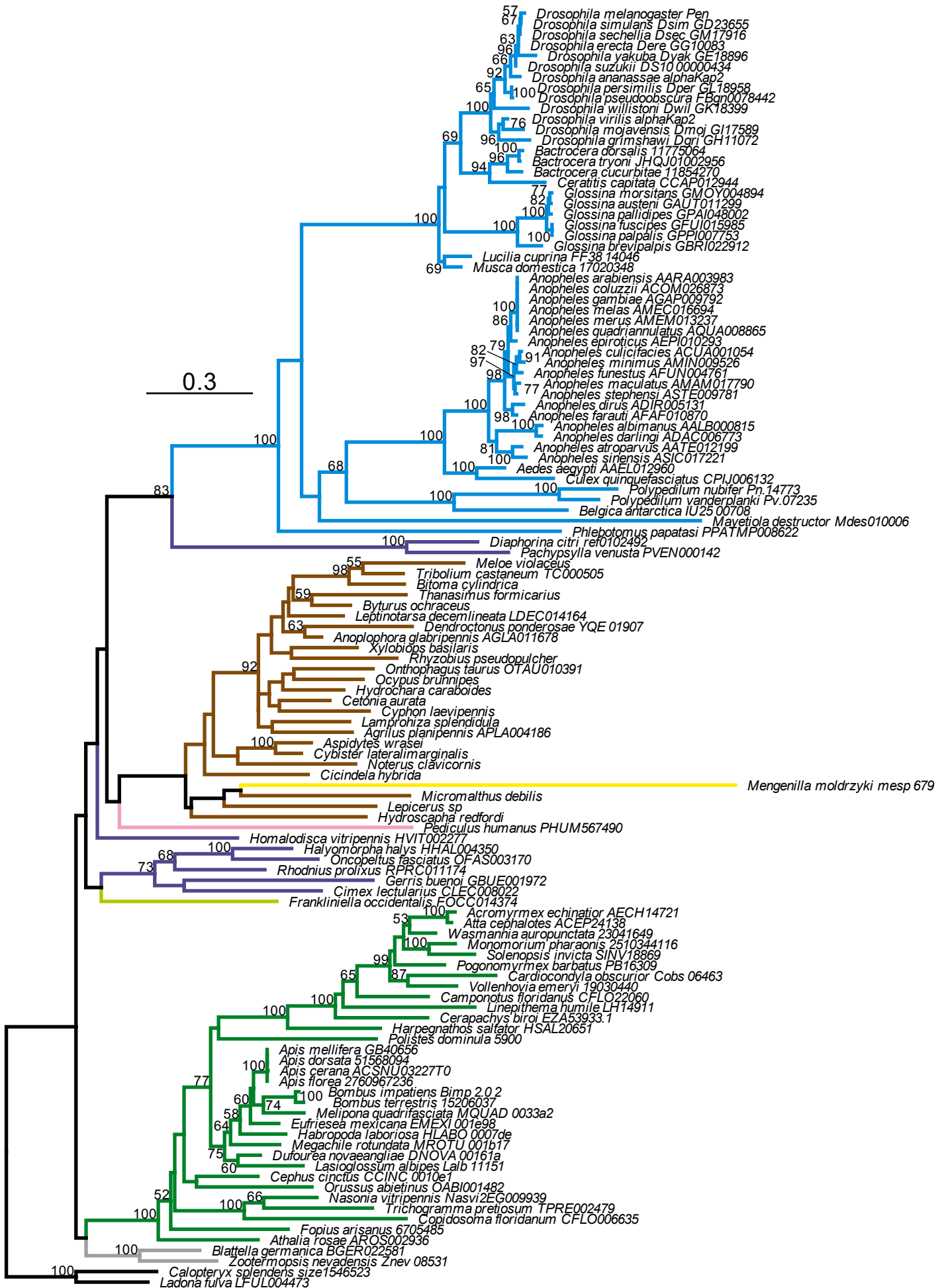
# oys



# Past1

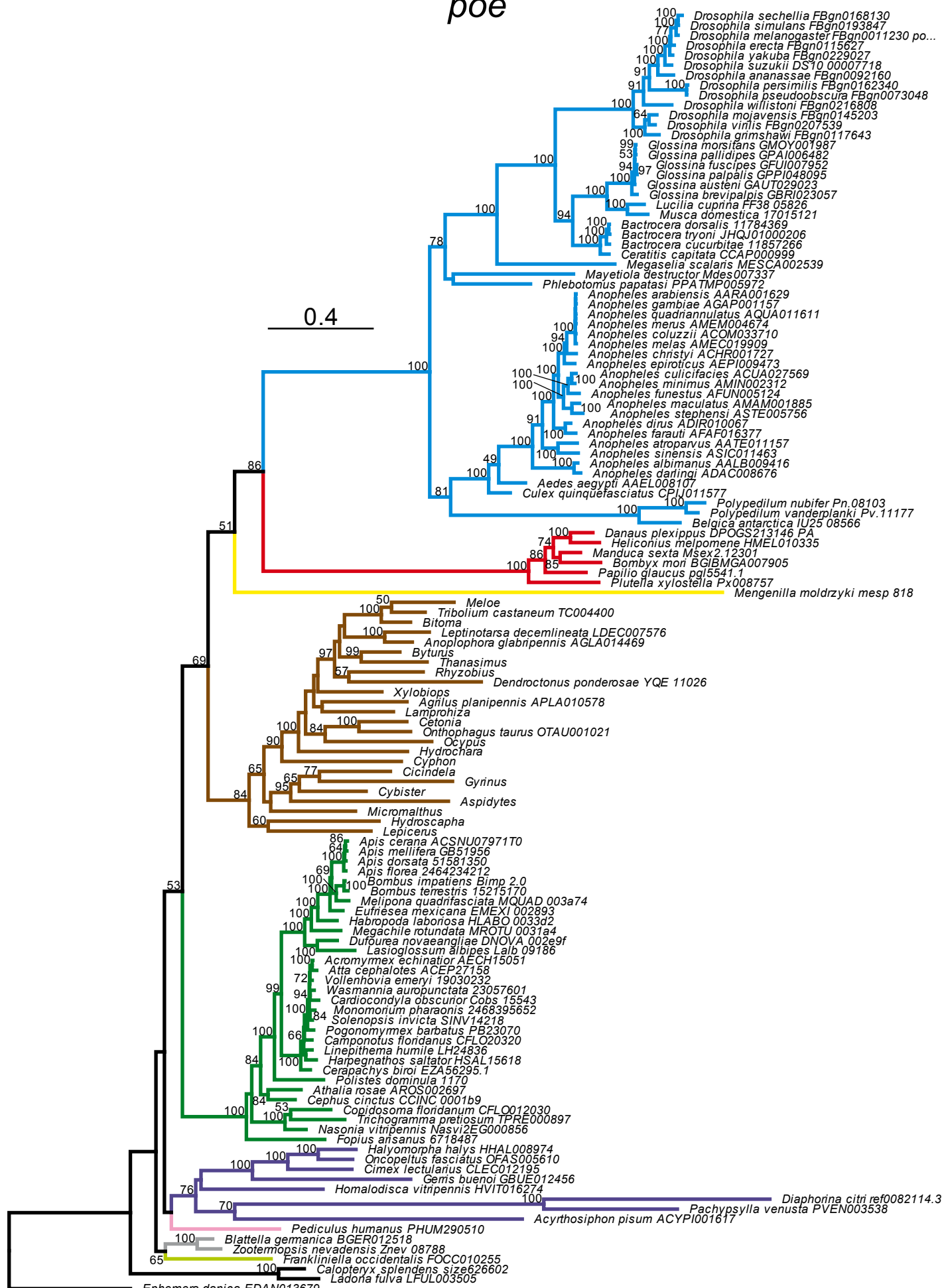


# Pen

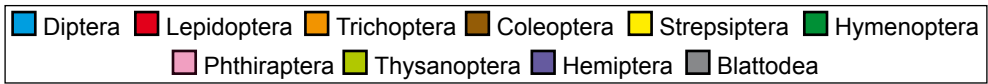
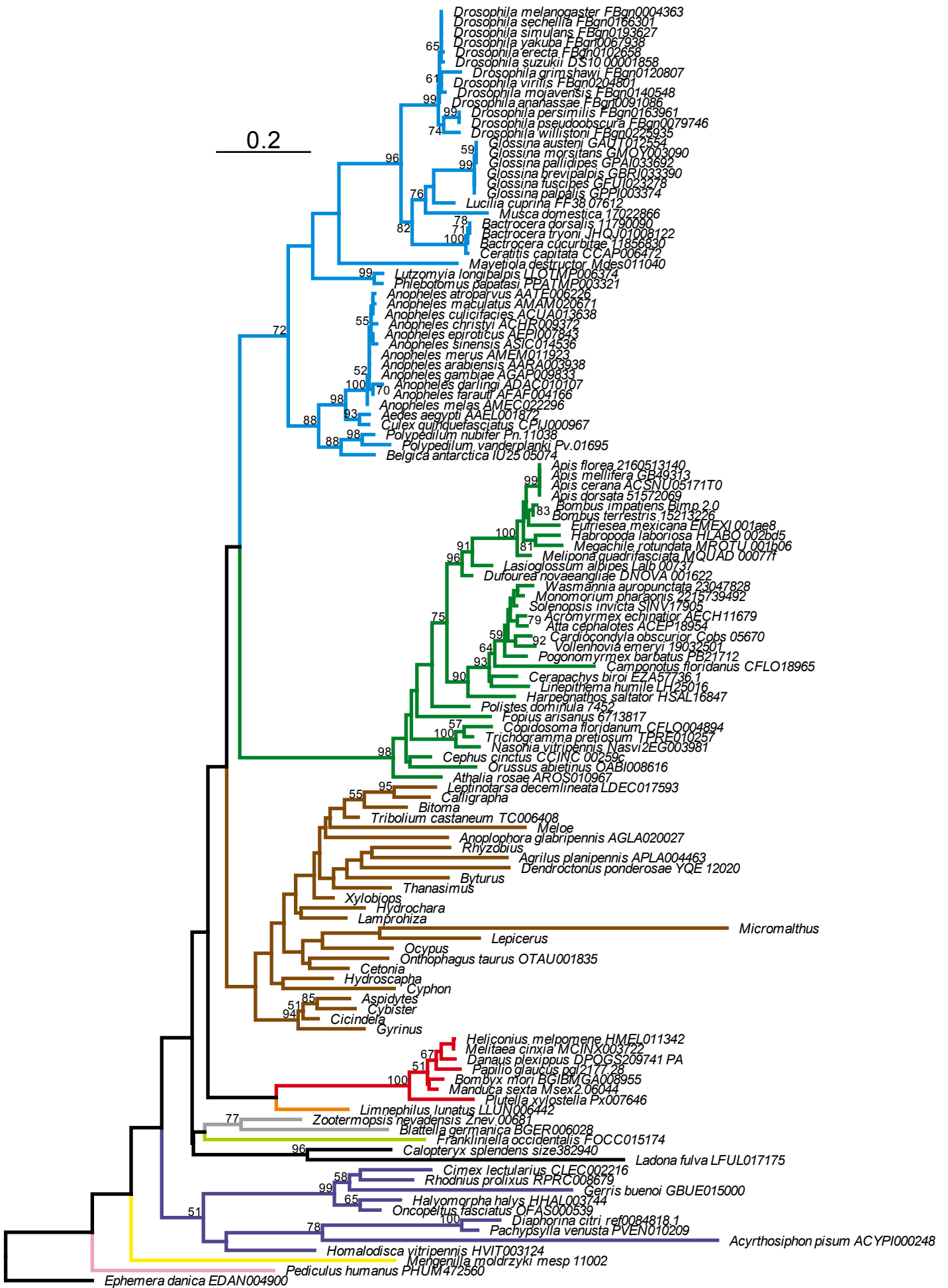




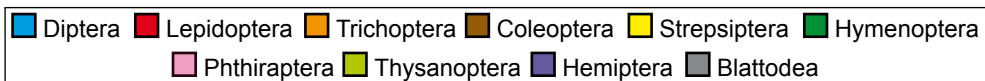
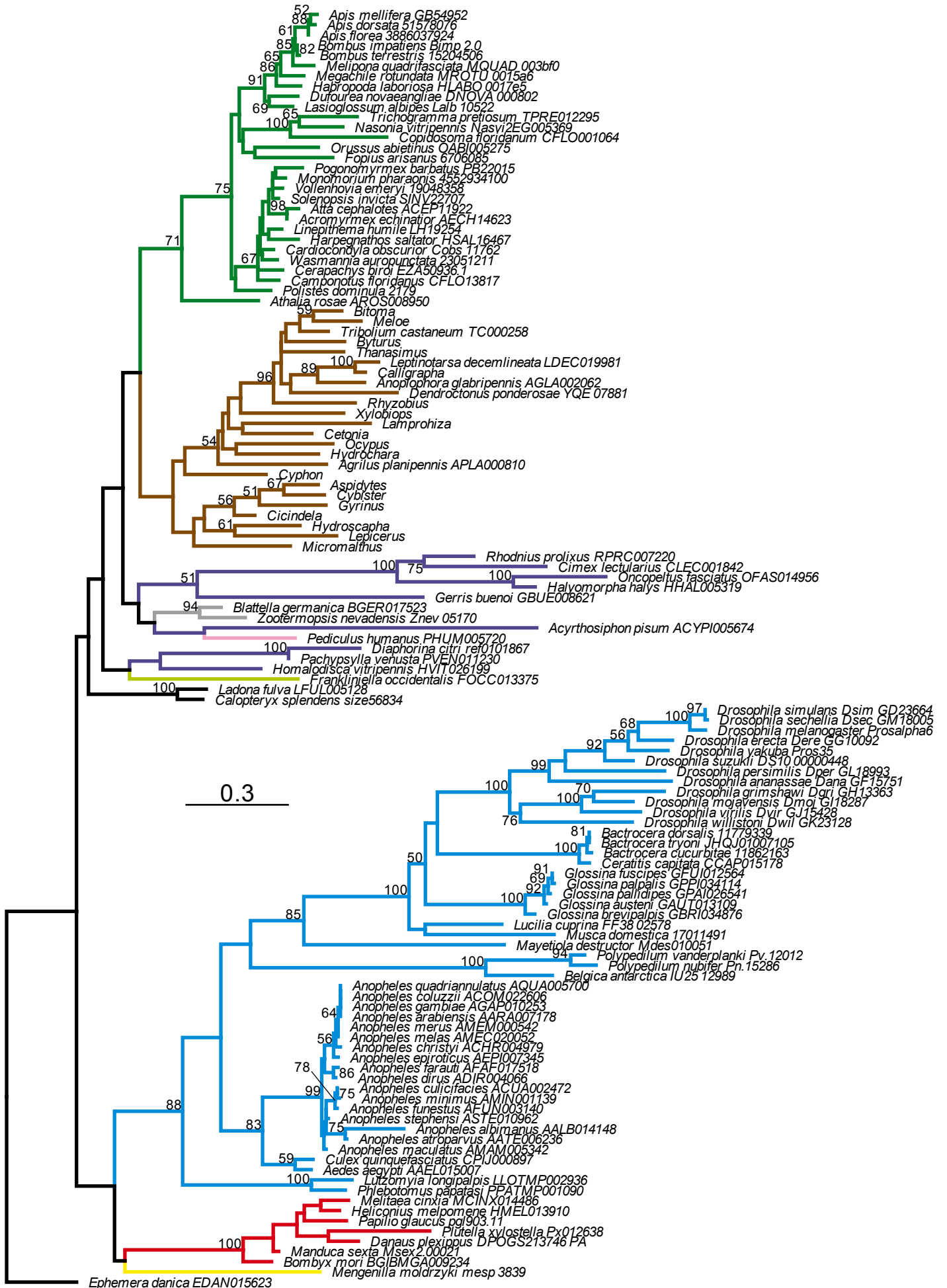
poe



# porin



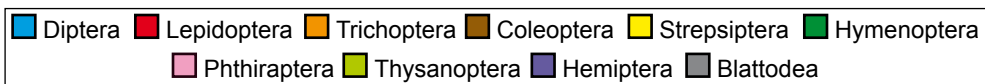
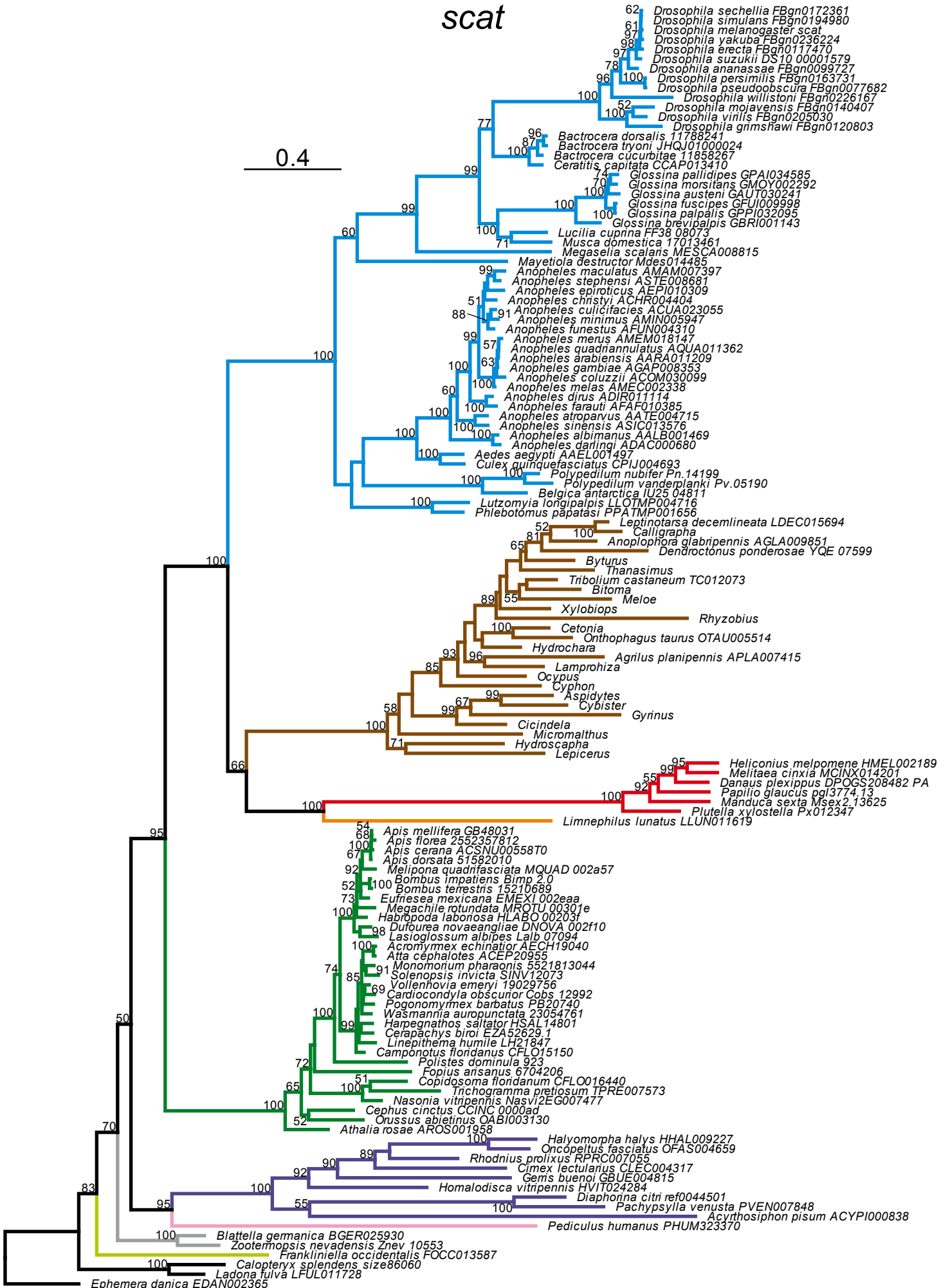
# Prosalpha6T



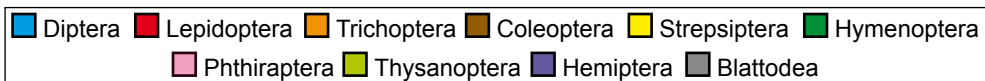
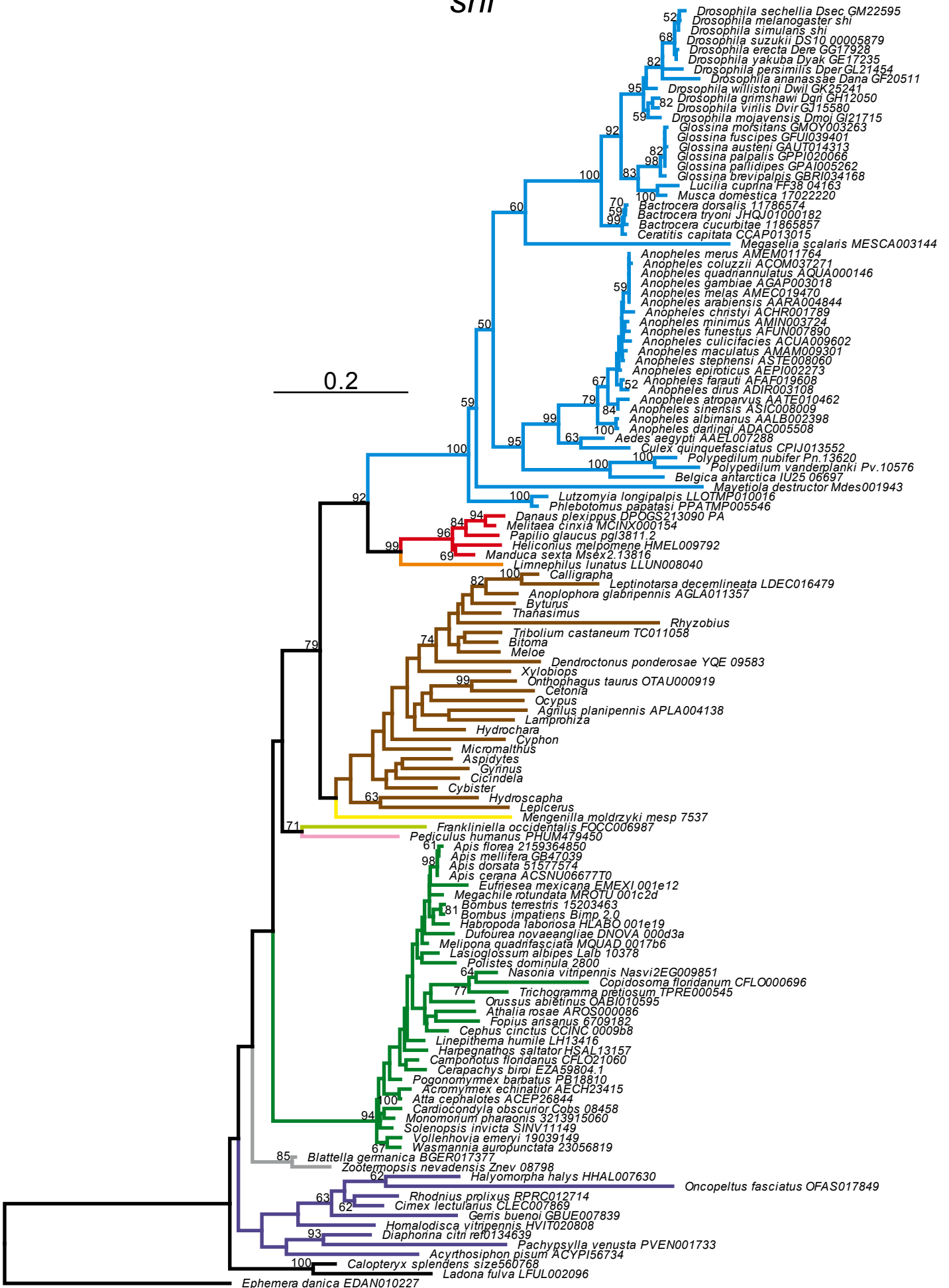


# scat

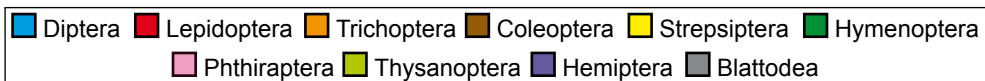
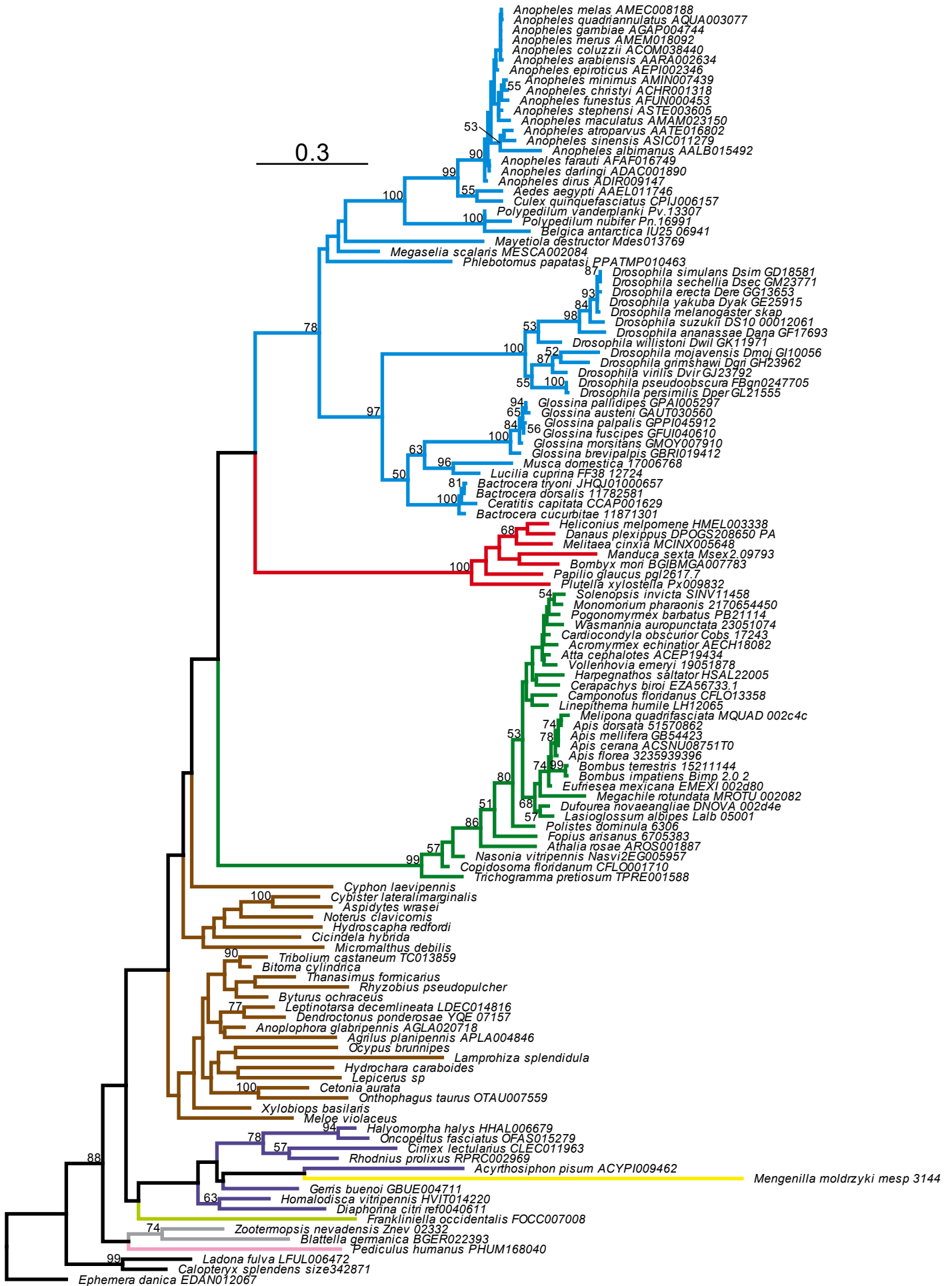
0.4



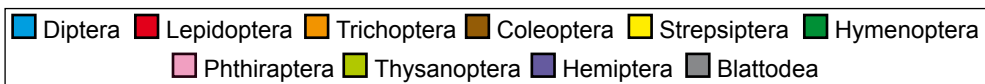
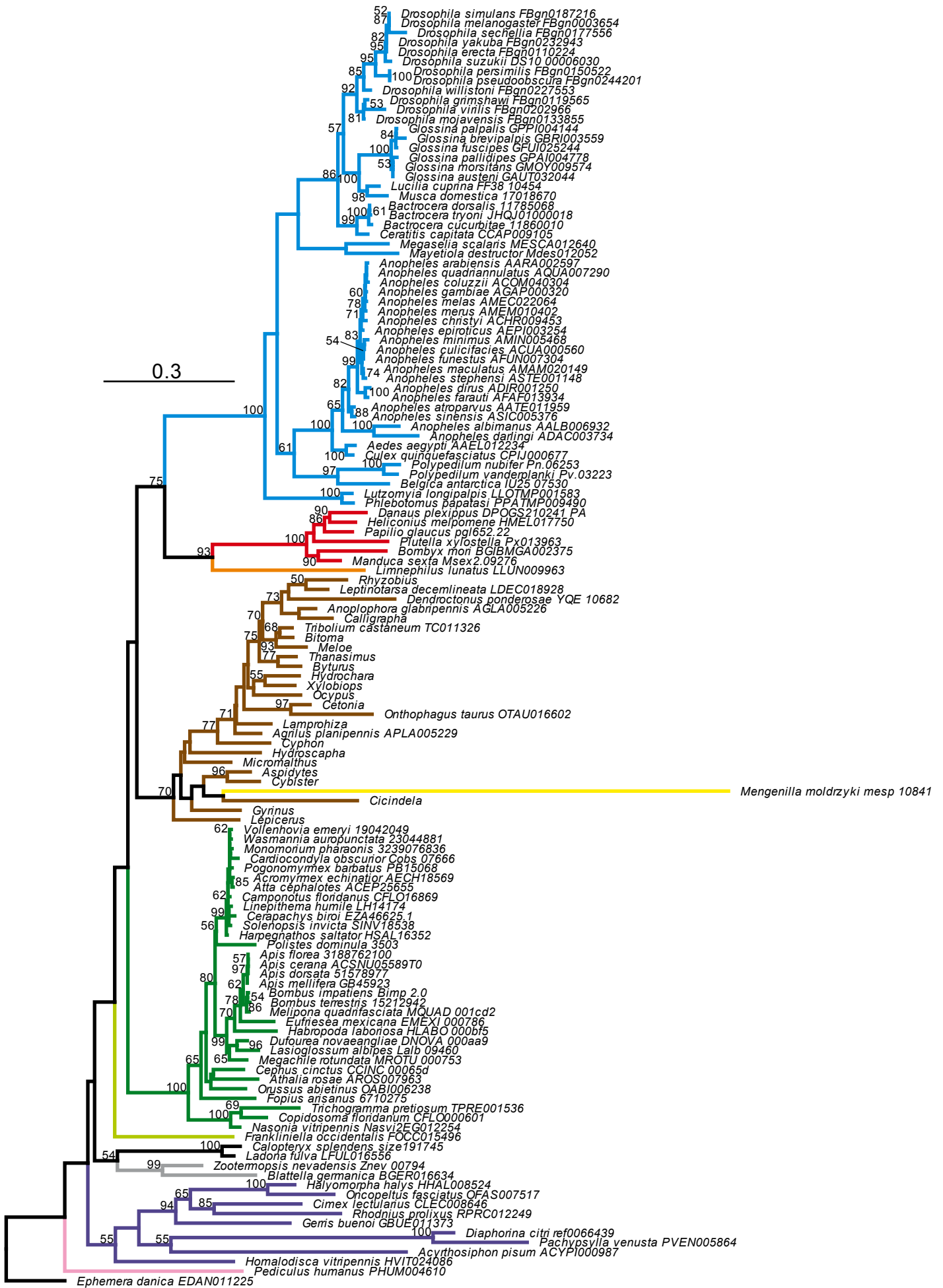
# shi



# skap

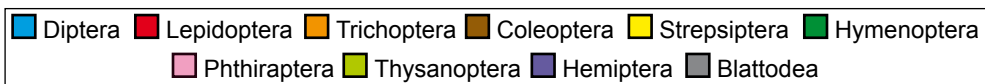
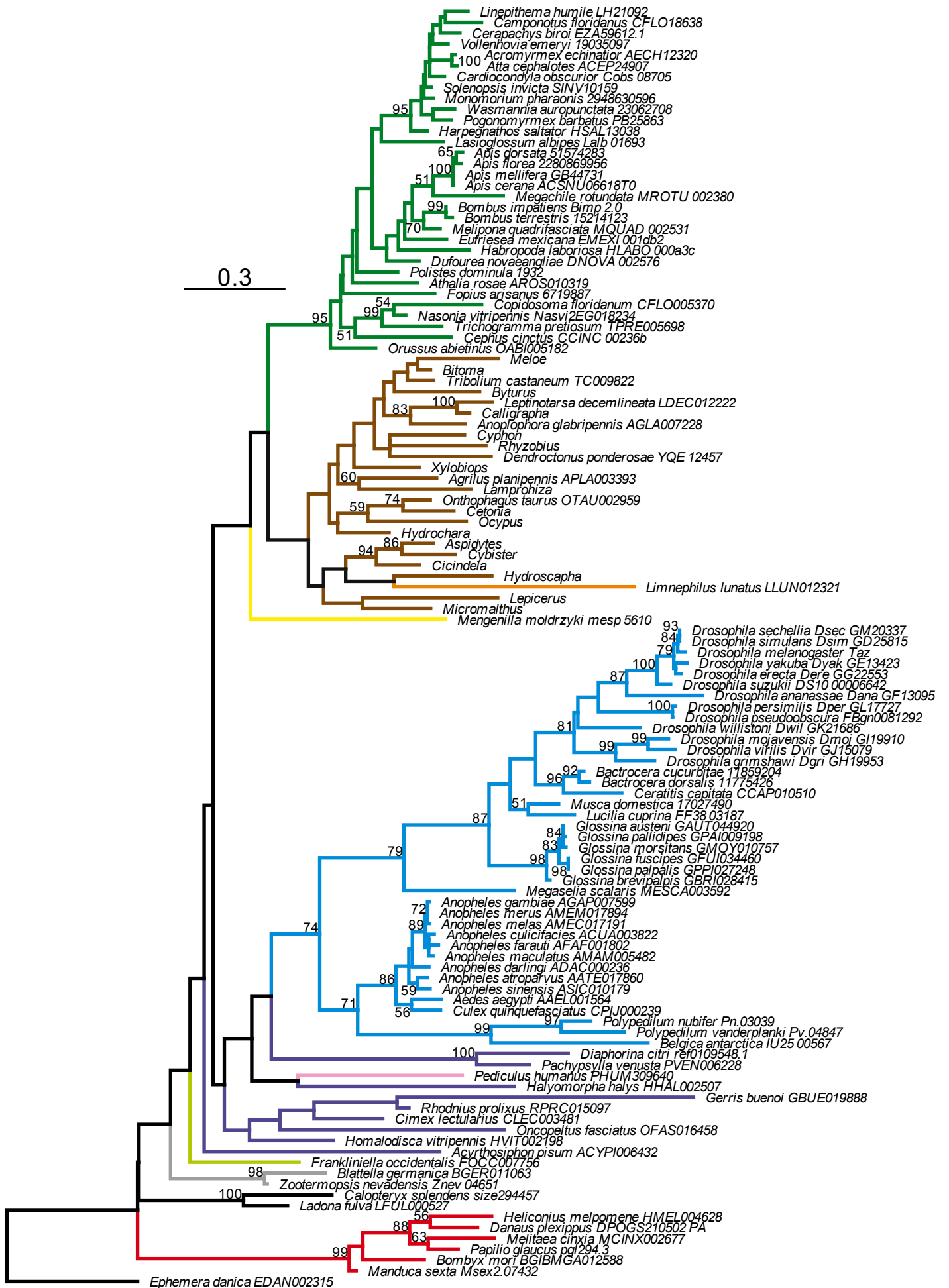


# SW

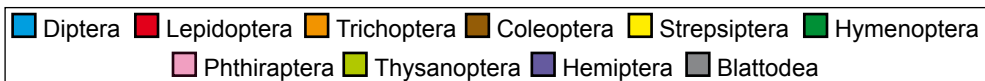
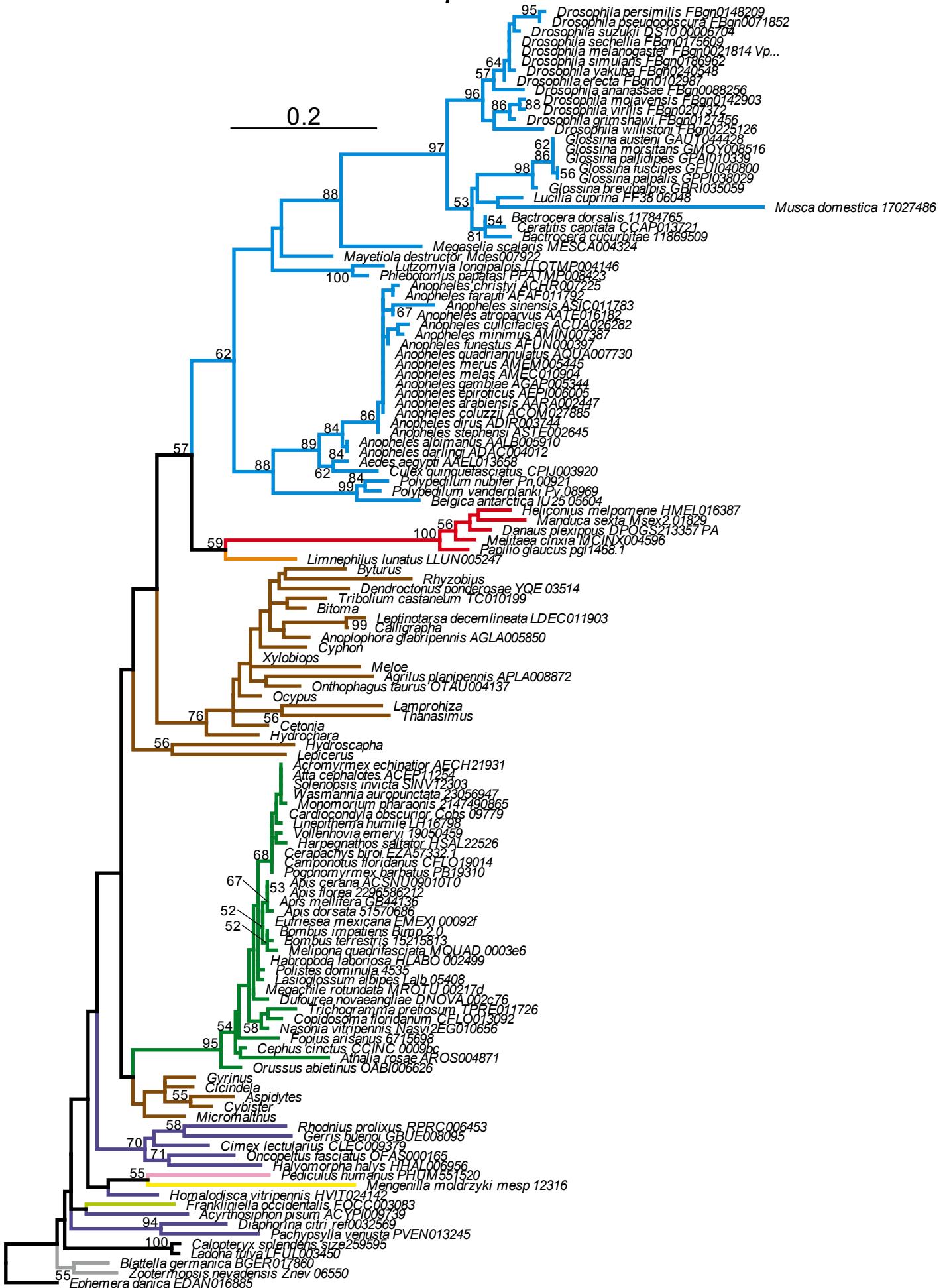




# Taz



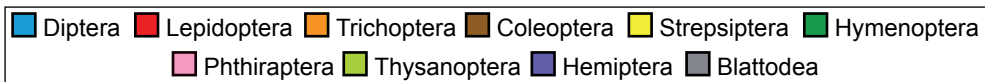
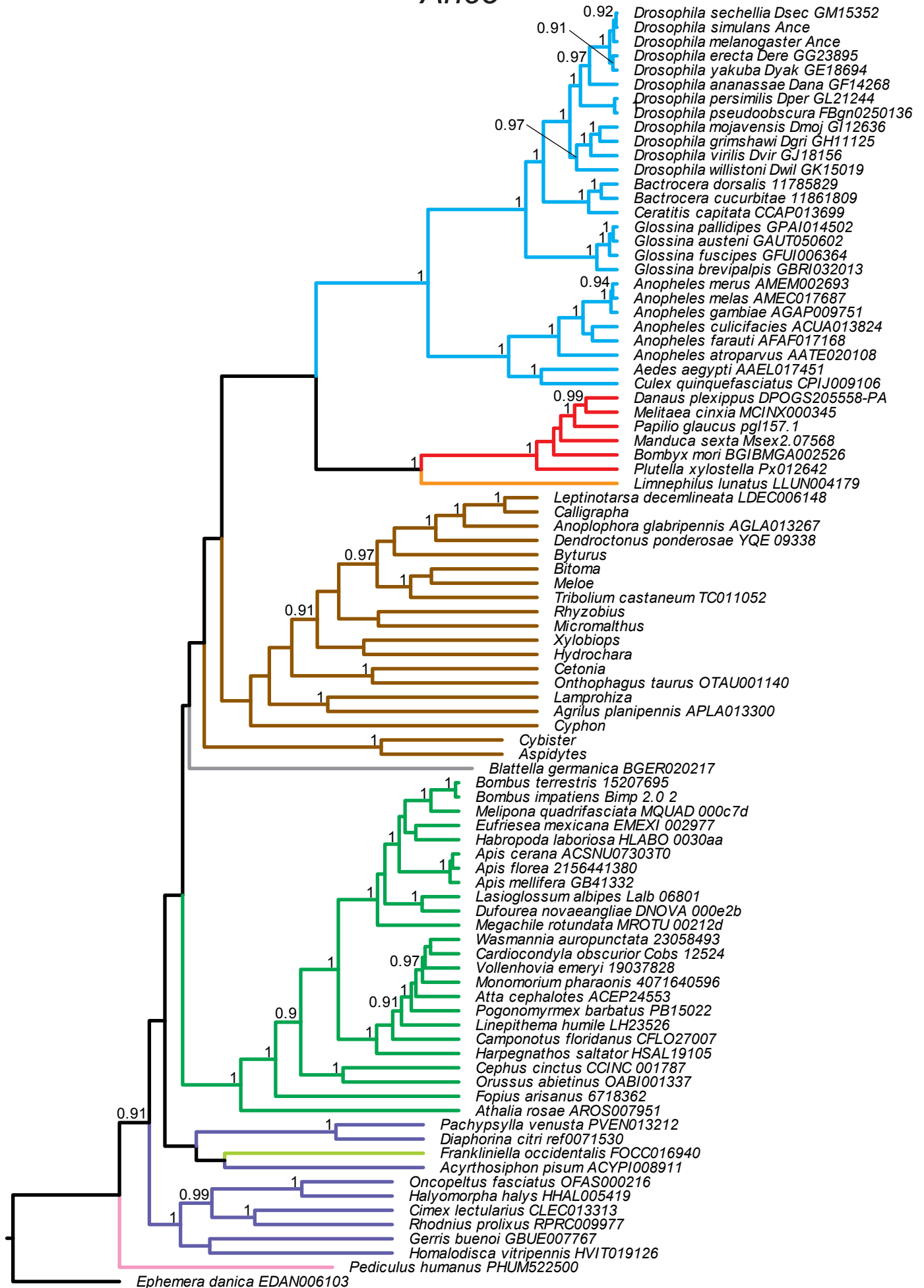
# Vps28



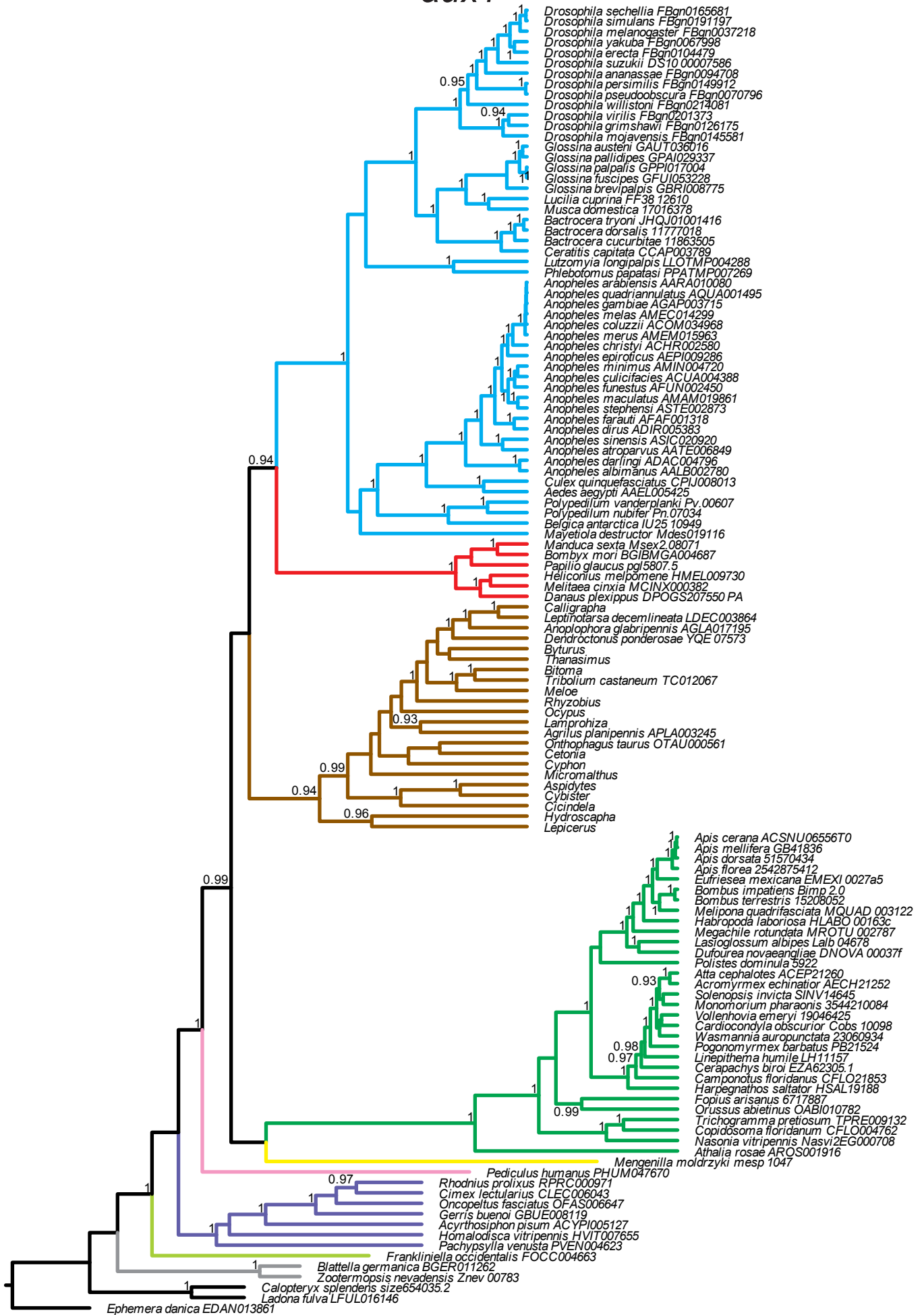
**File S2.** Bayesian inference trees based on the amino acid alignments of different sperm individualization proteins in insects.



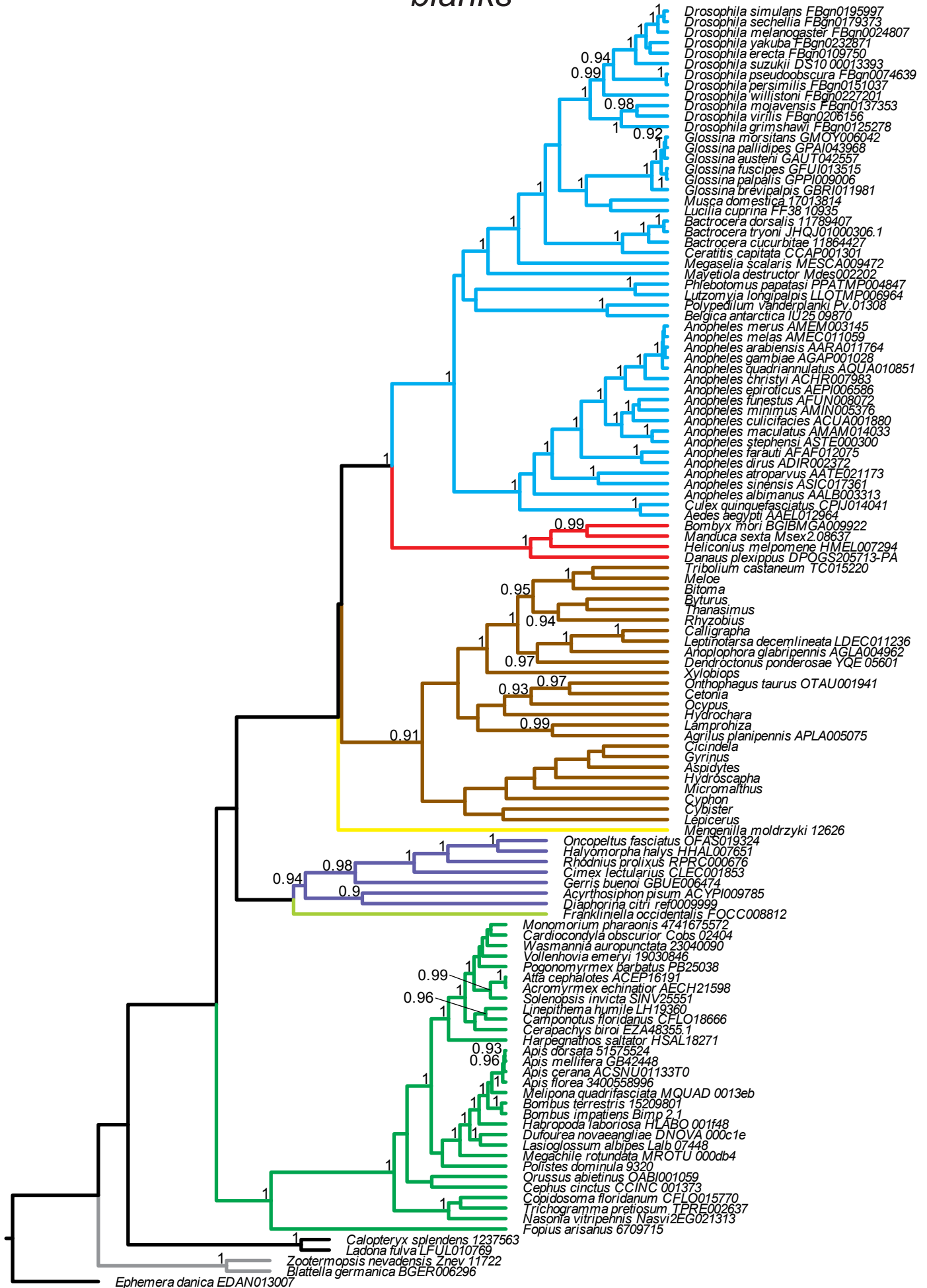
# Ance



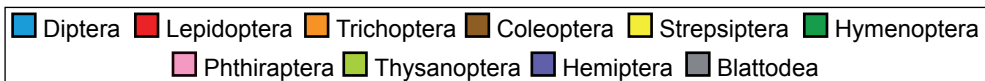
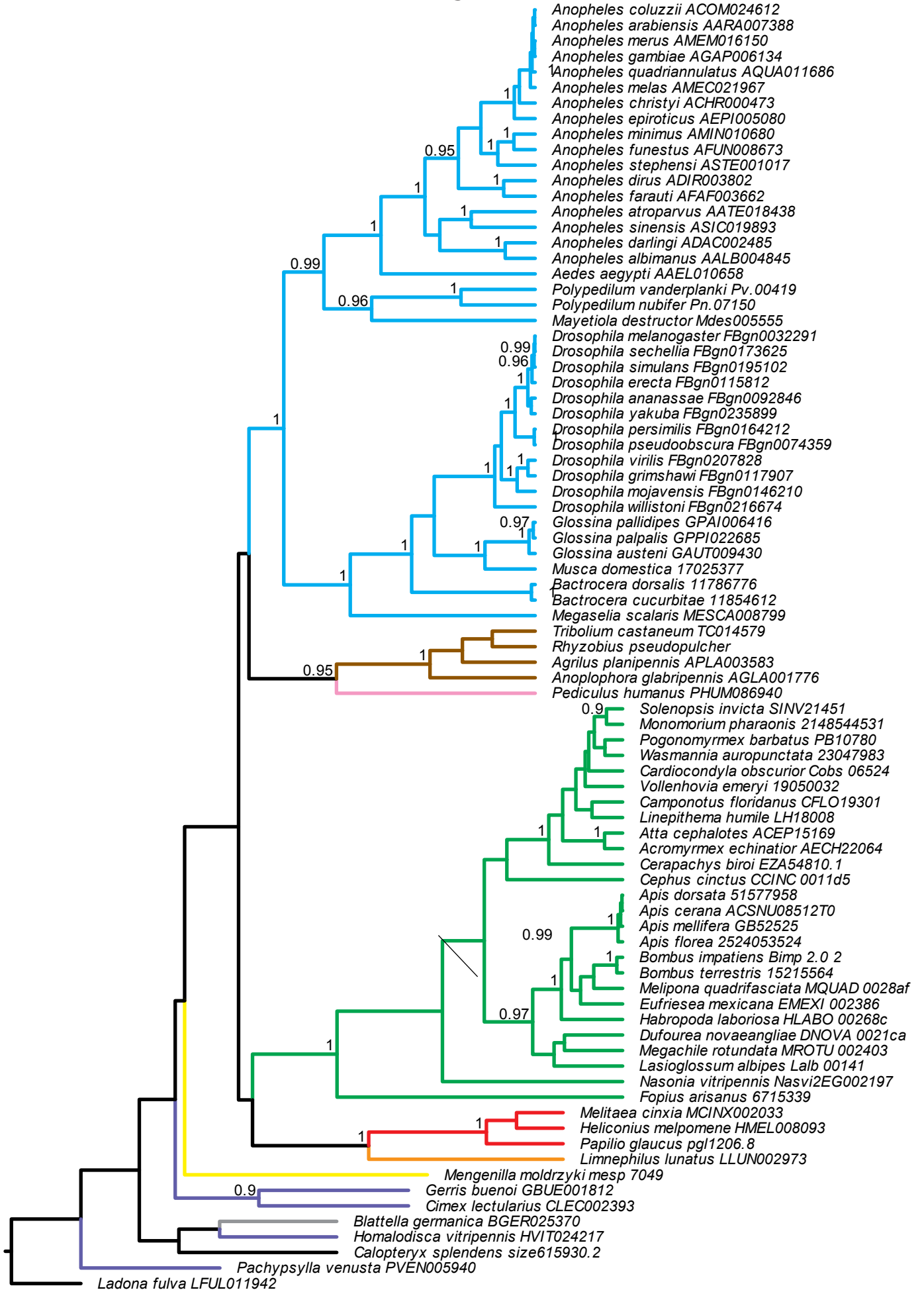
# aux1



# blanks

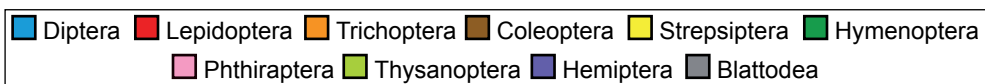
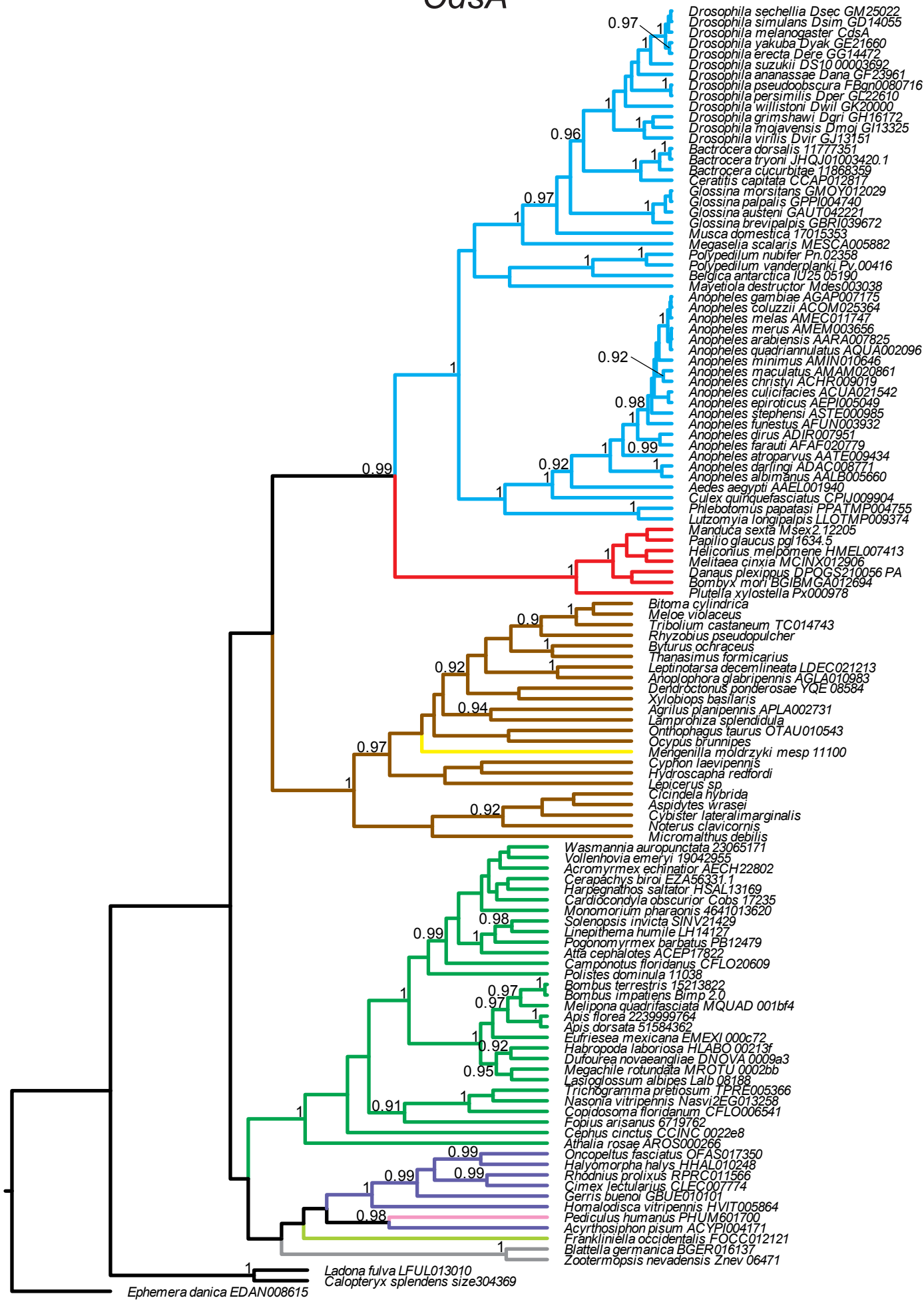


# Bug22

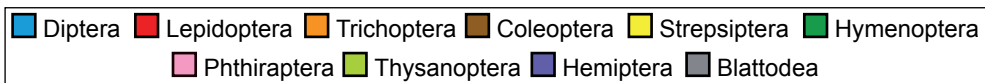
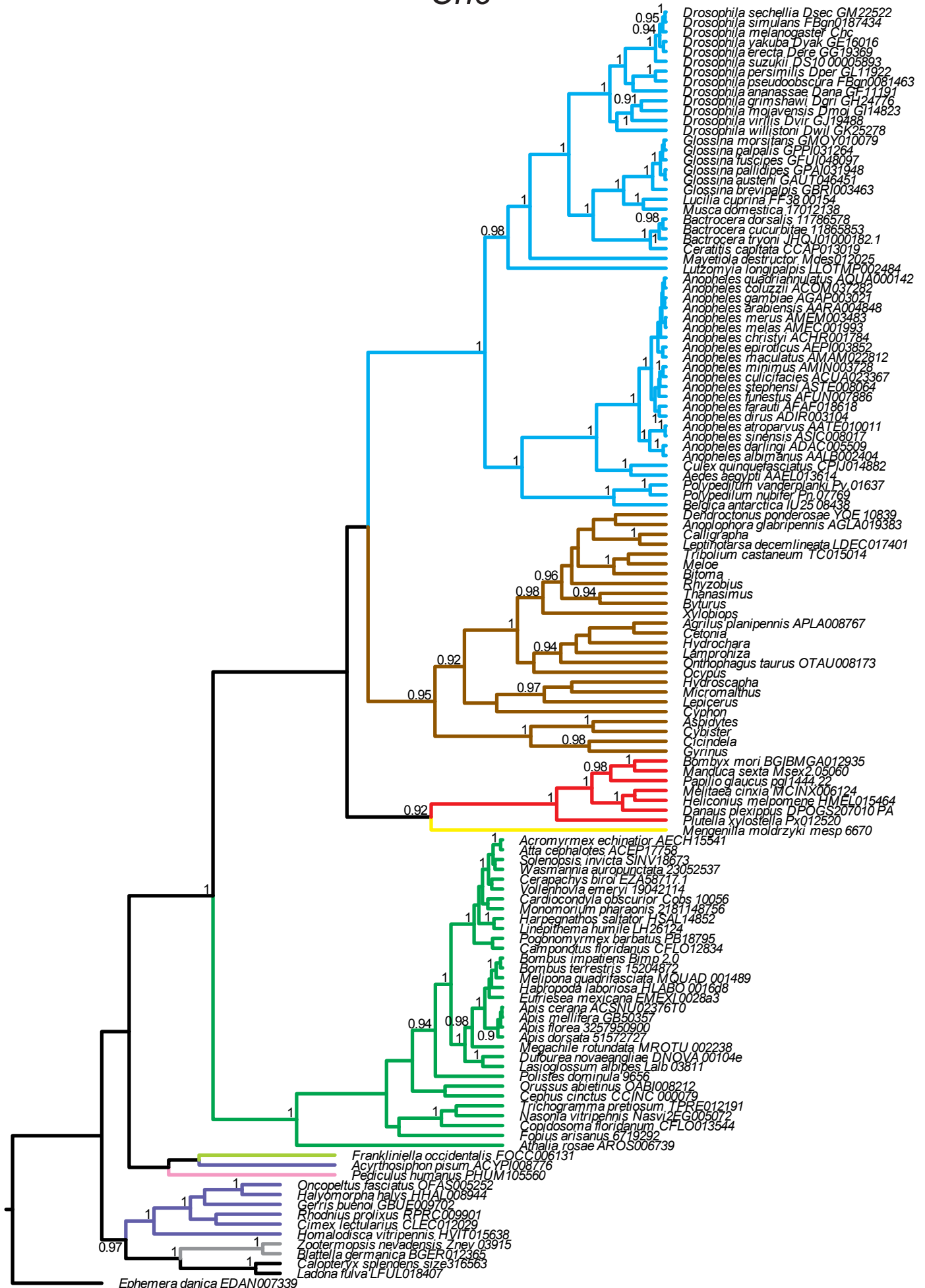




# CdsA

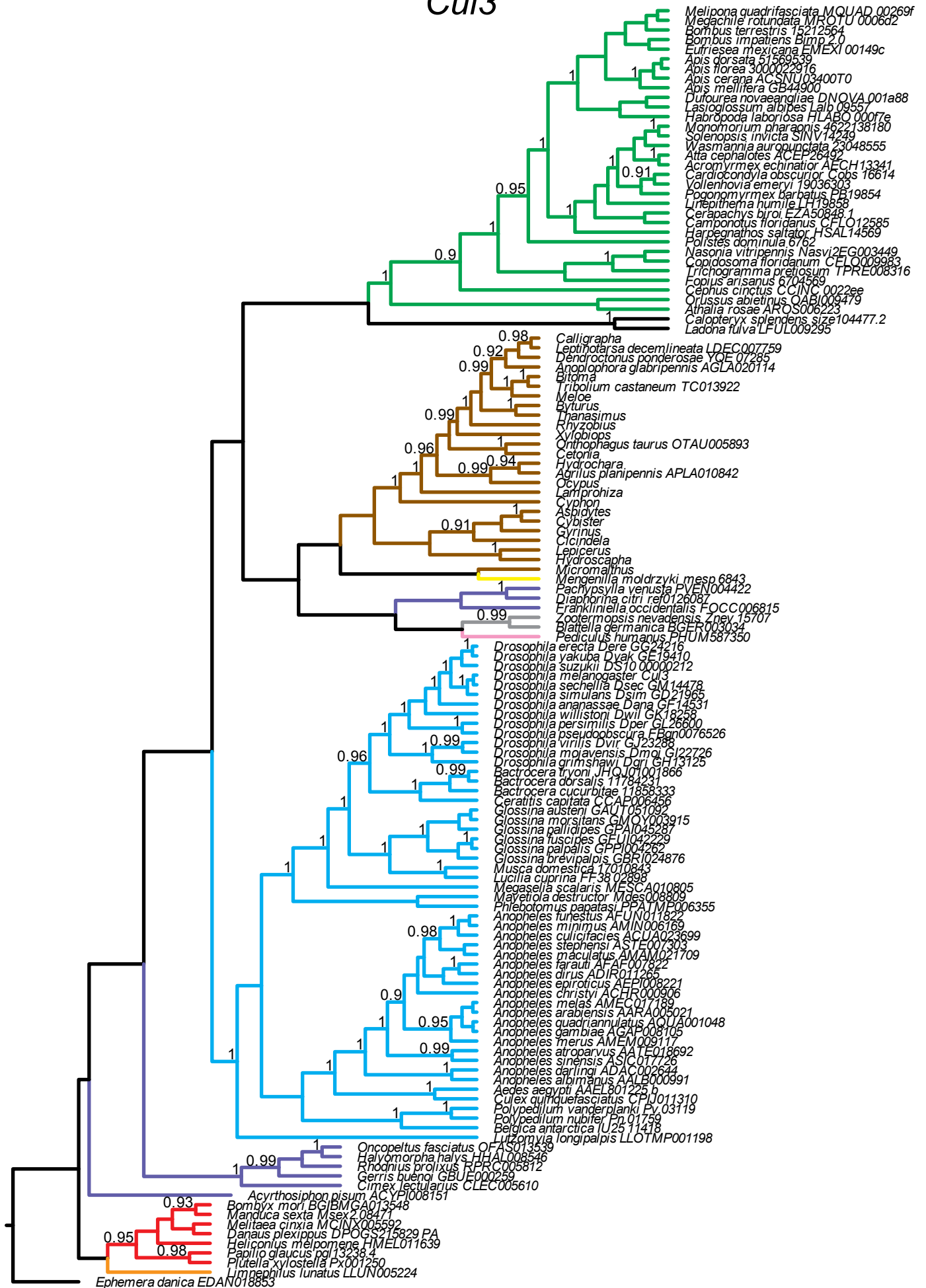


# Chc

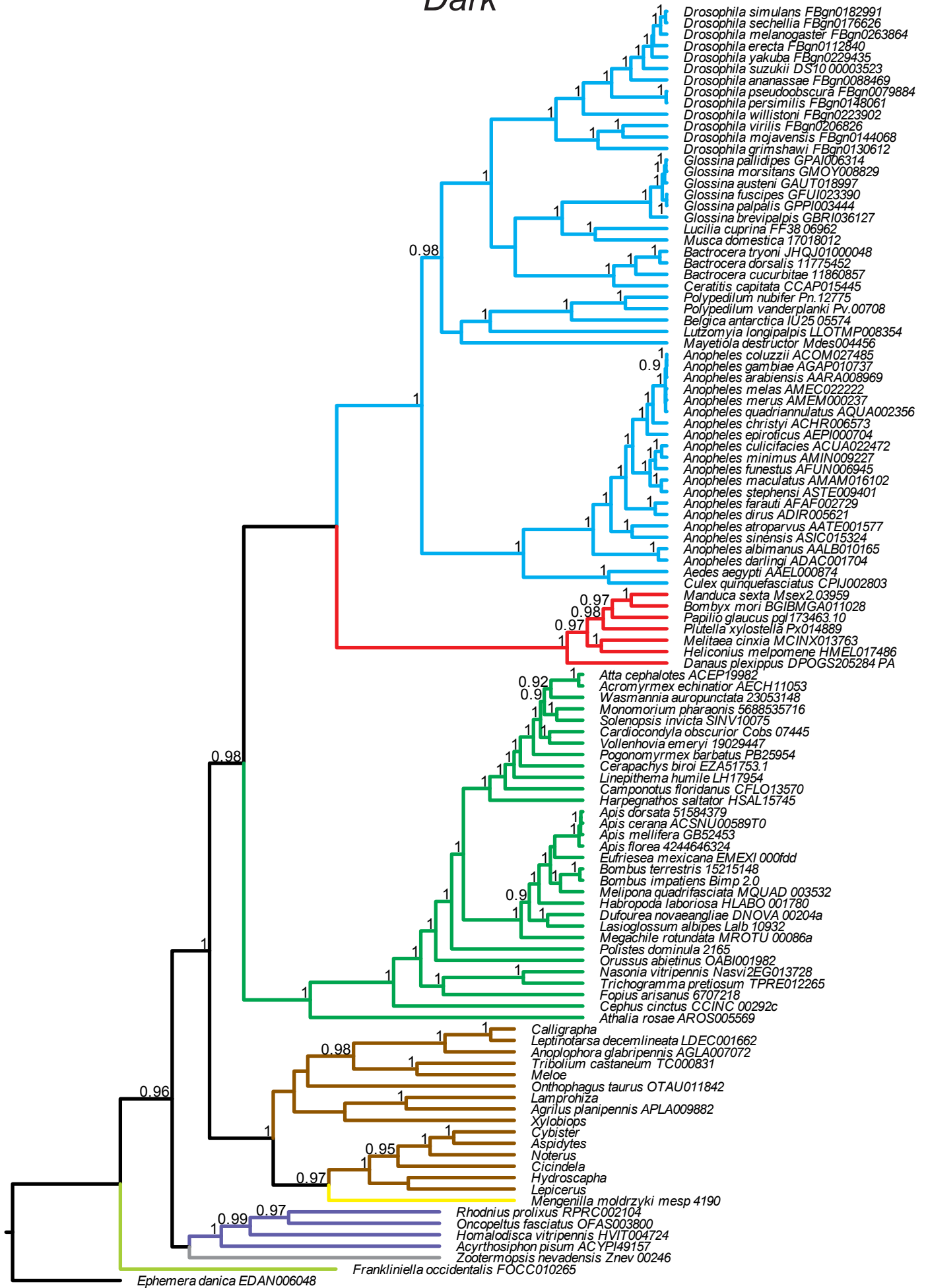




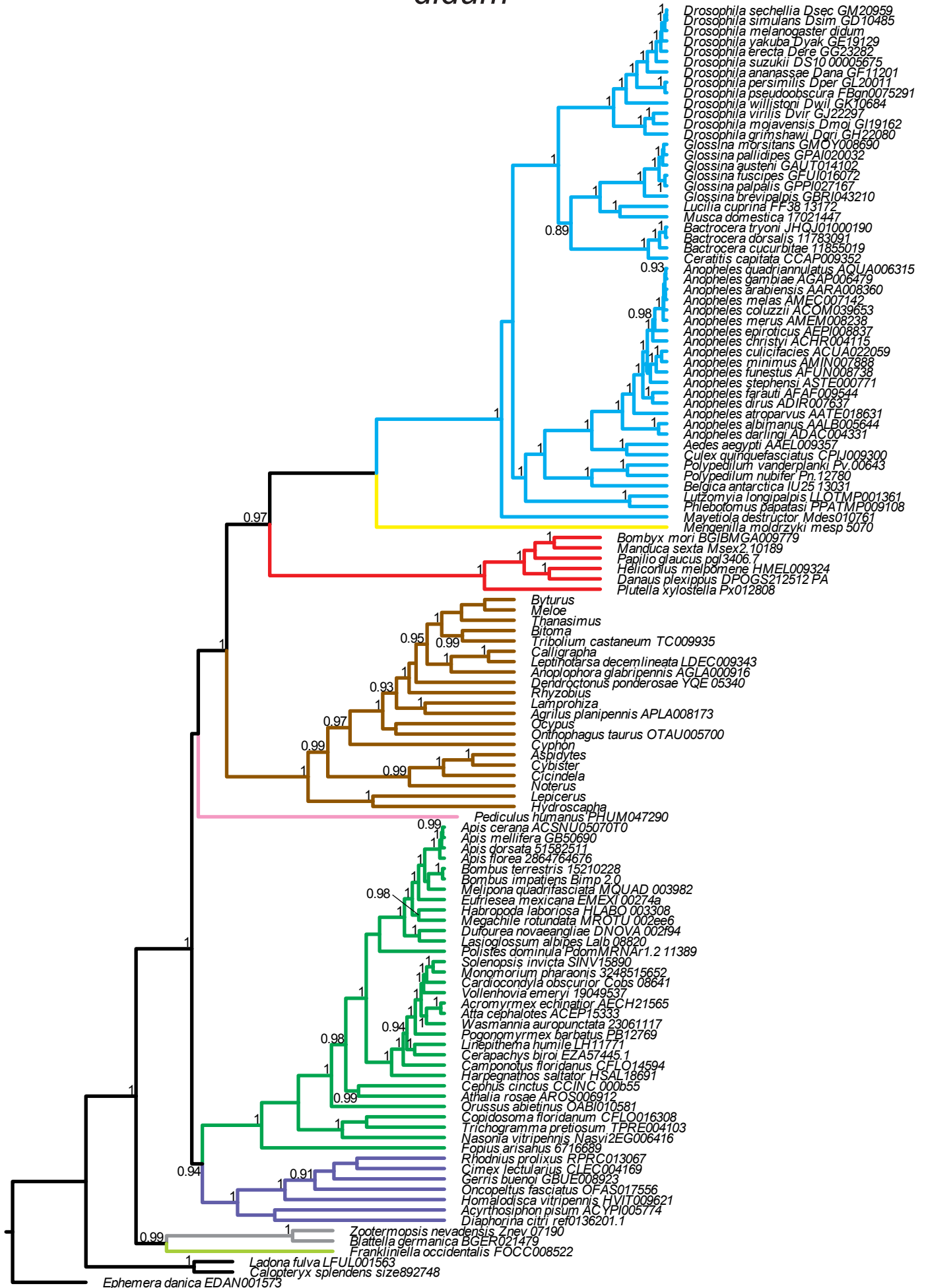
# Cul3



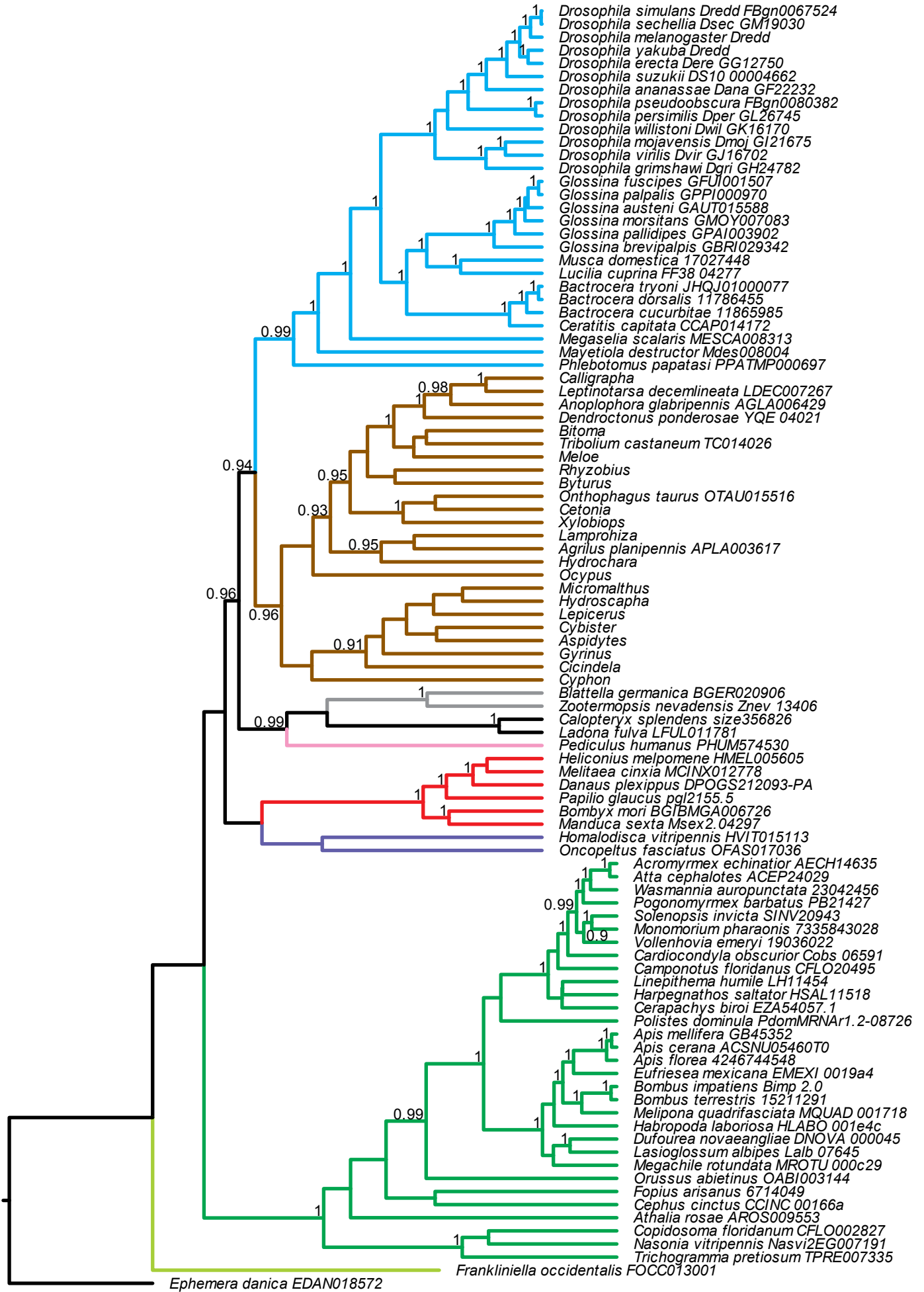
# Dark



# didum

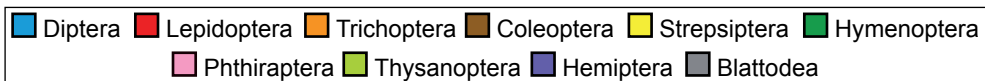
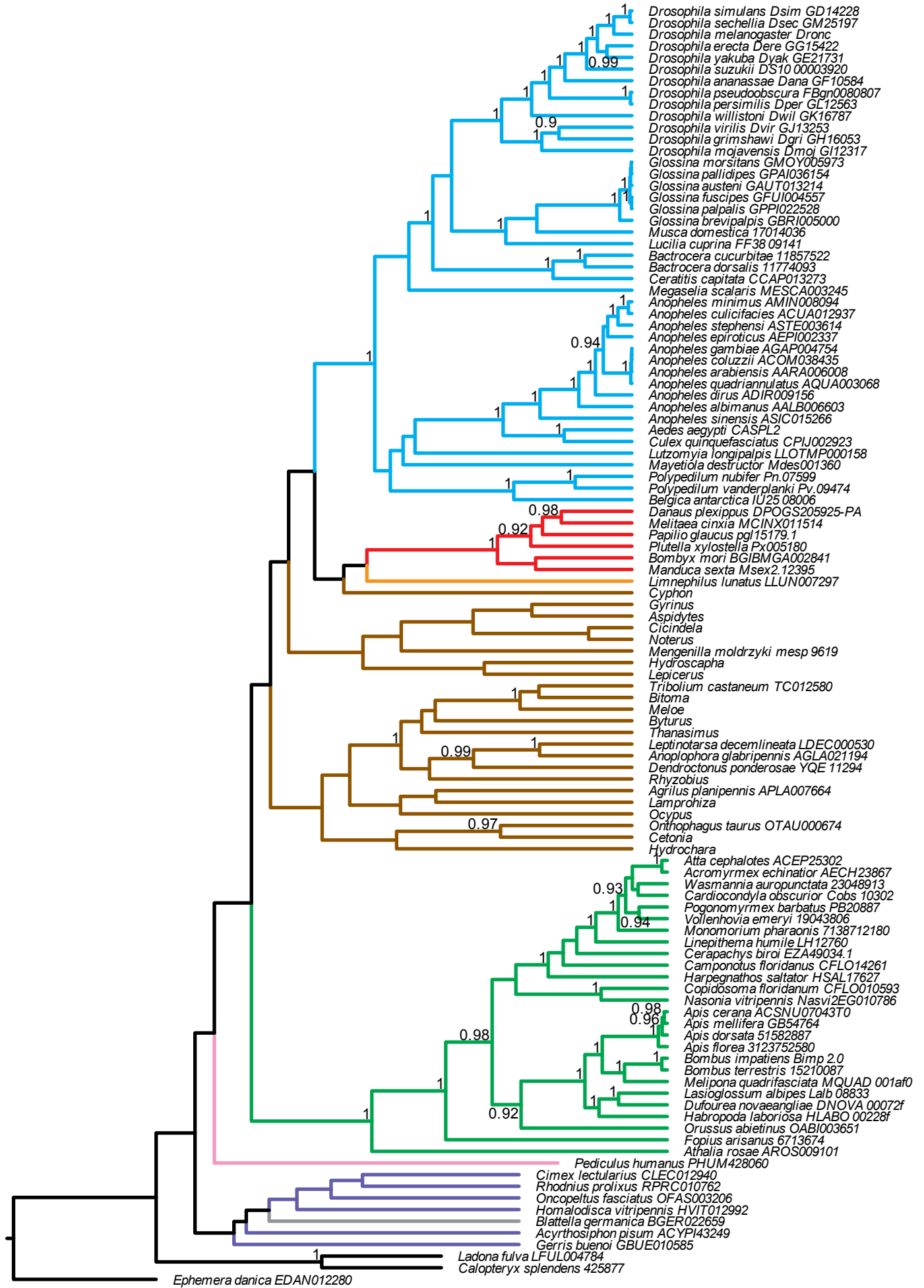


# Dredd

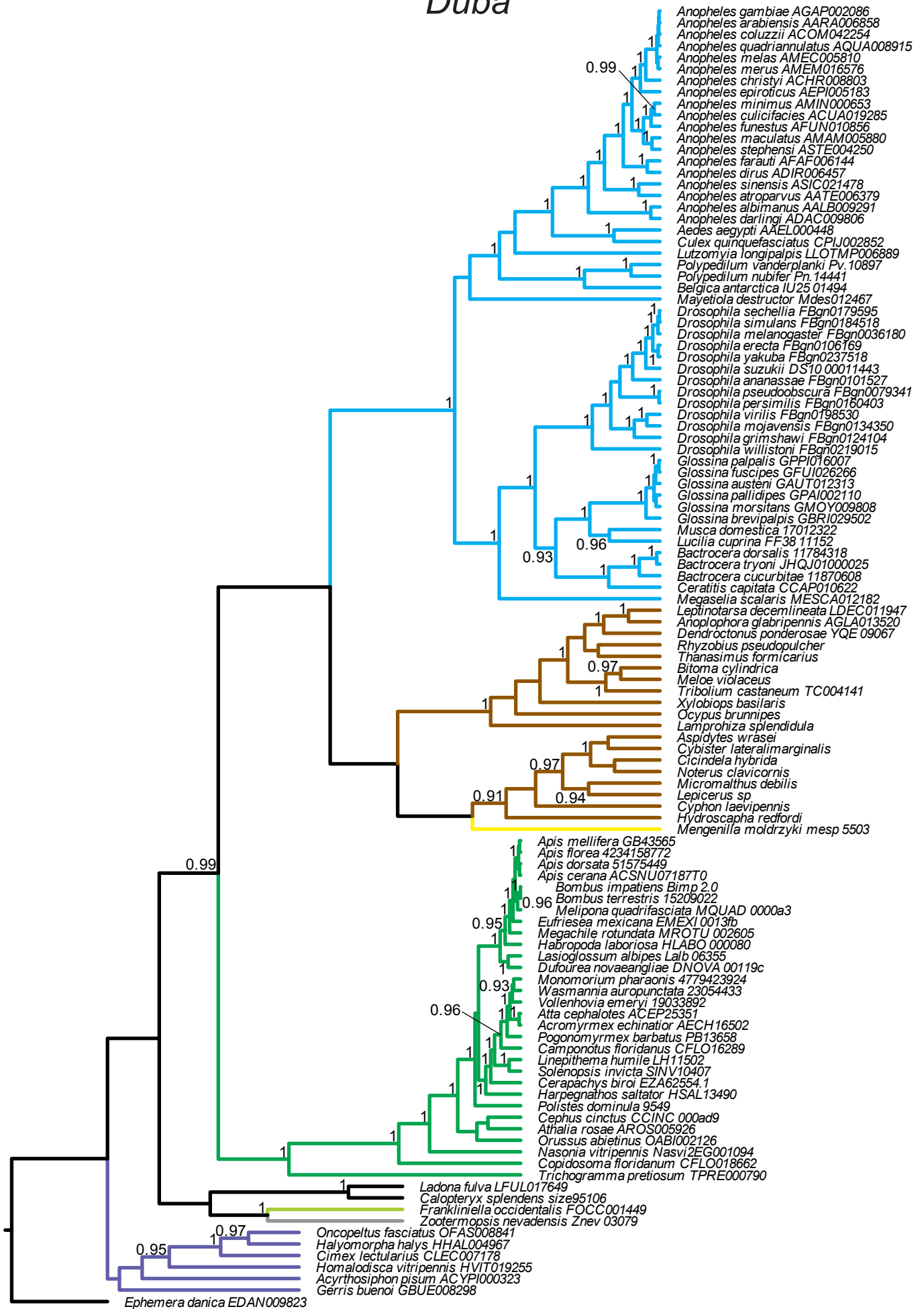




# Dronc

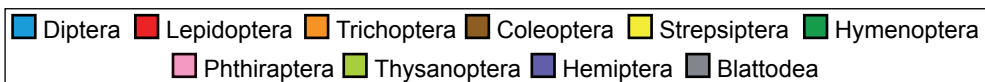
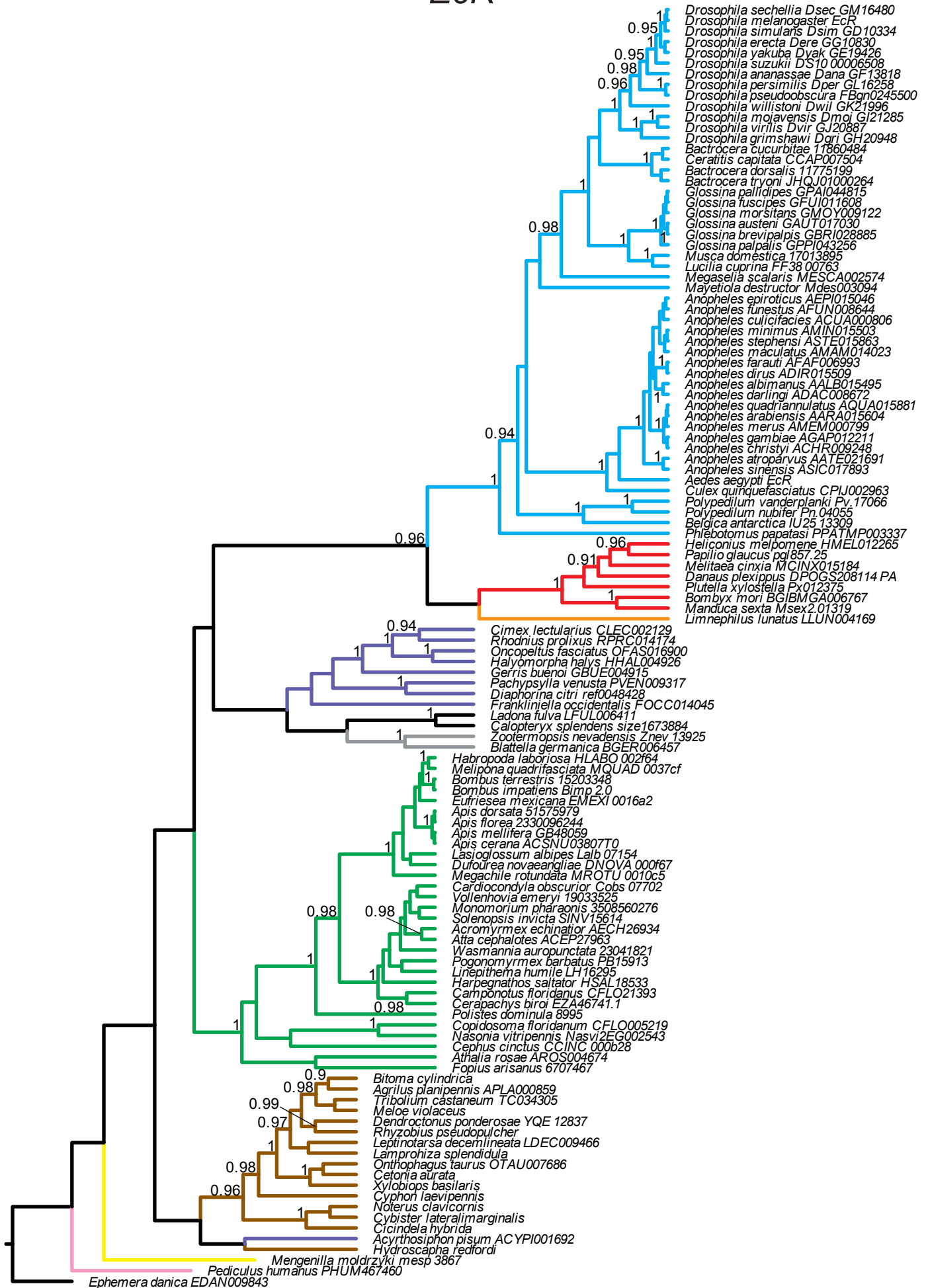


# Duba

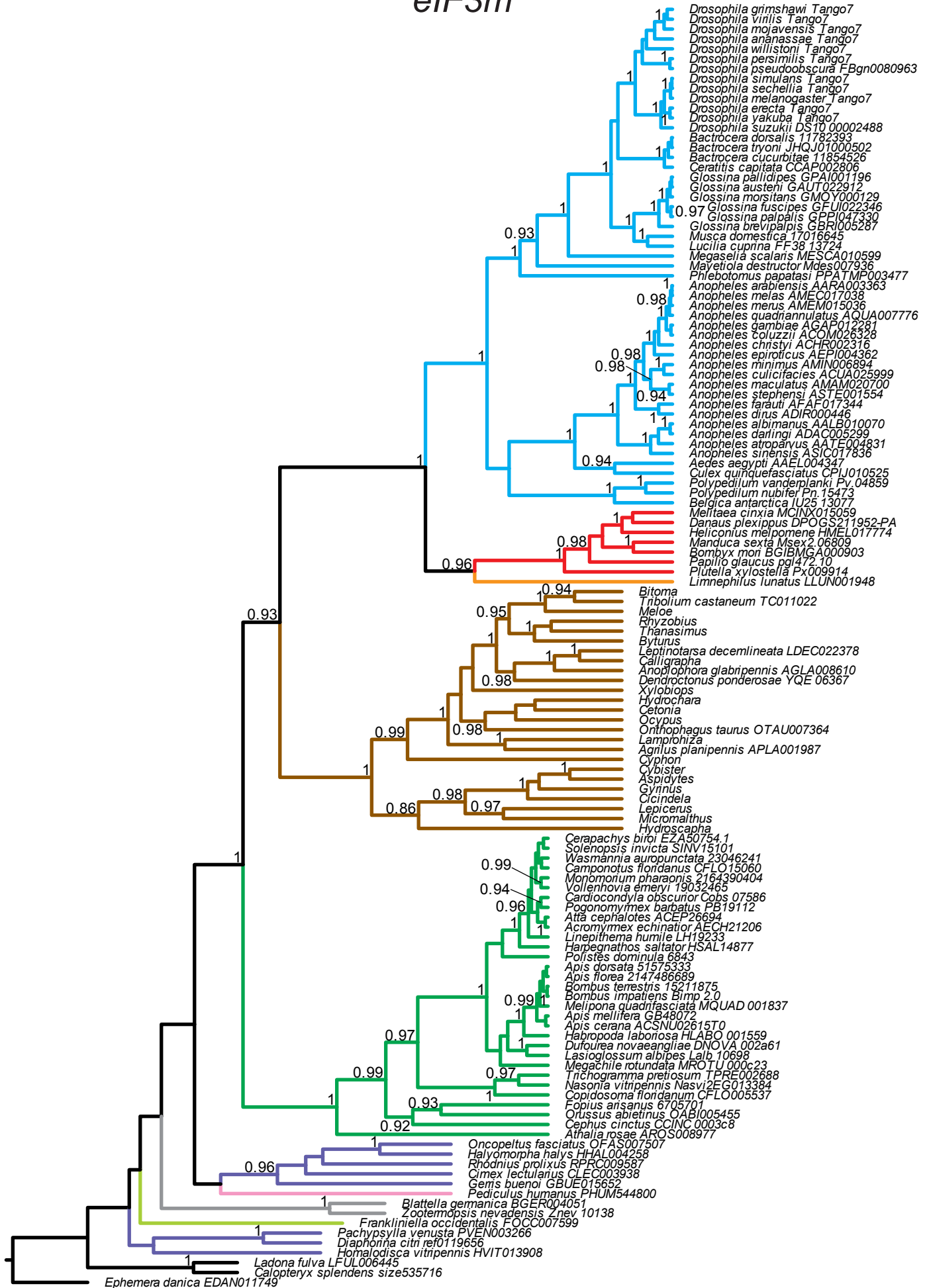




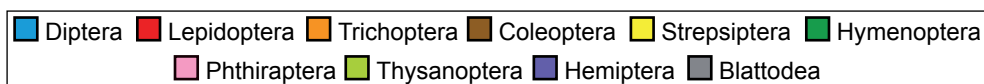
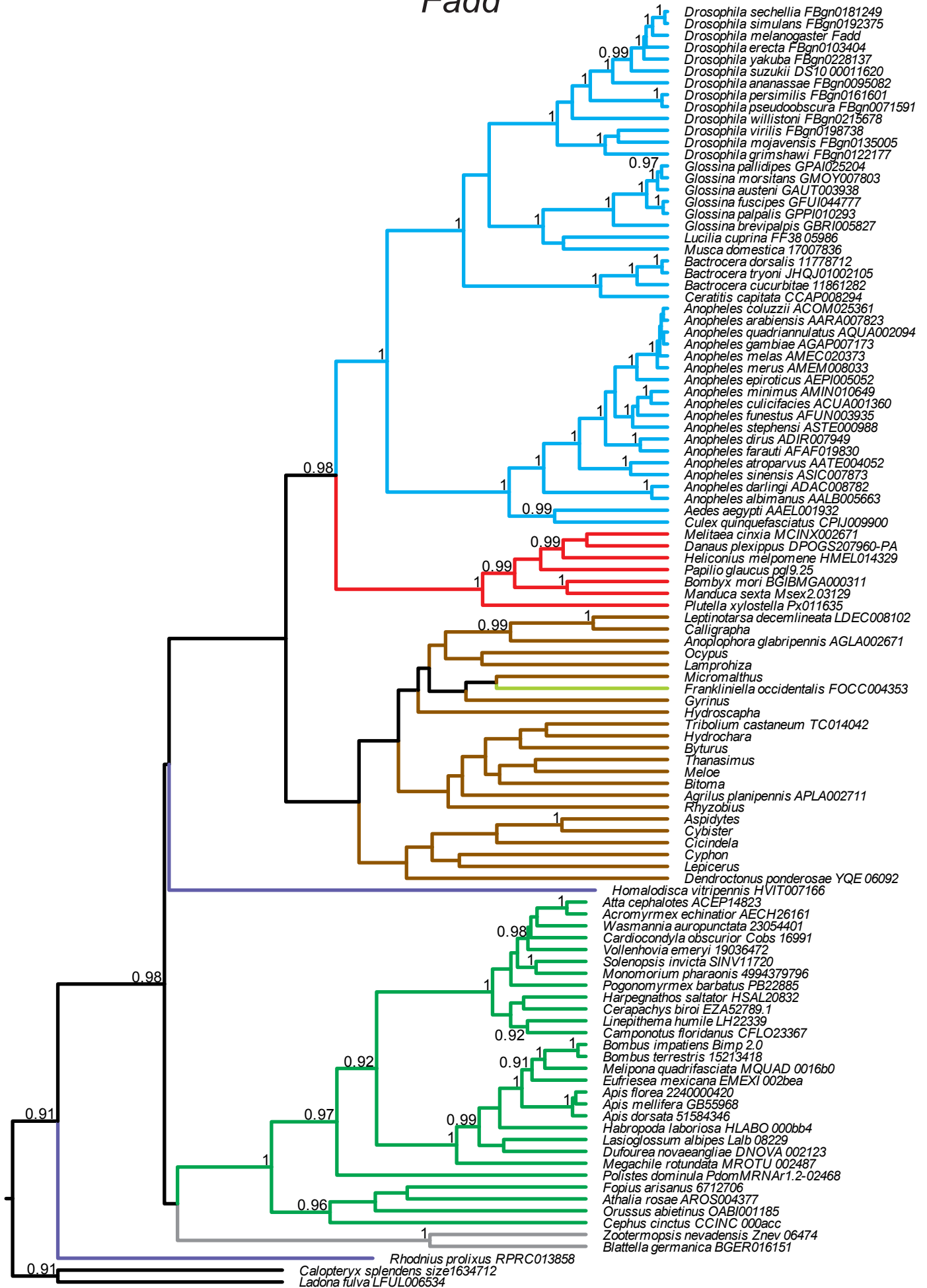
# EcR



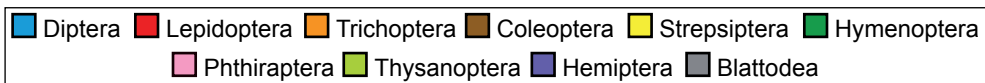
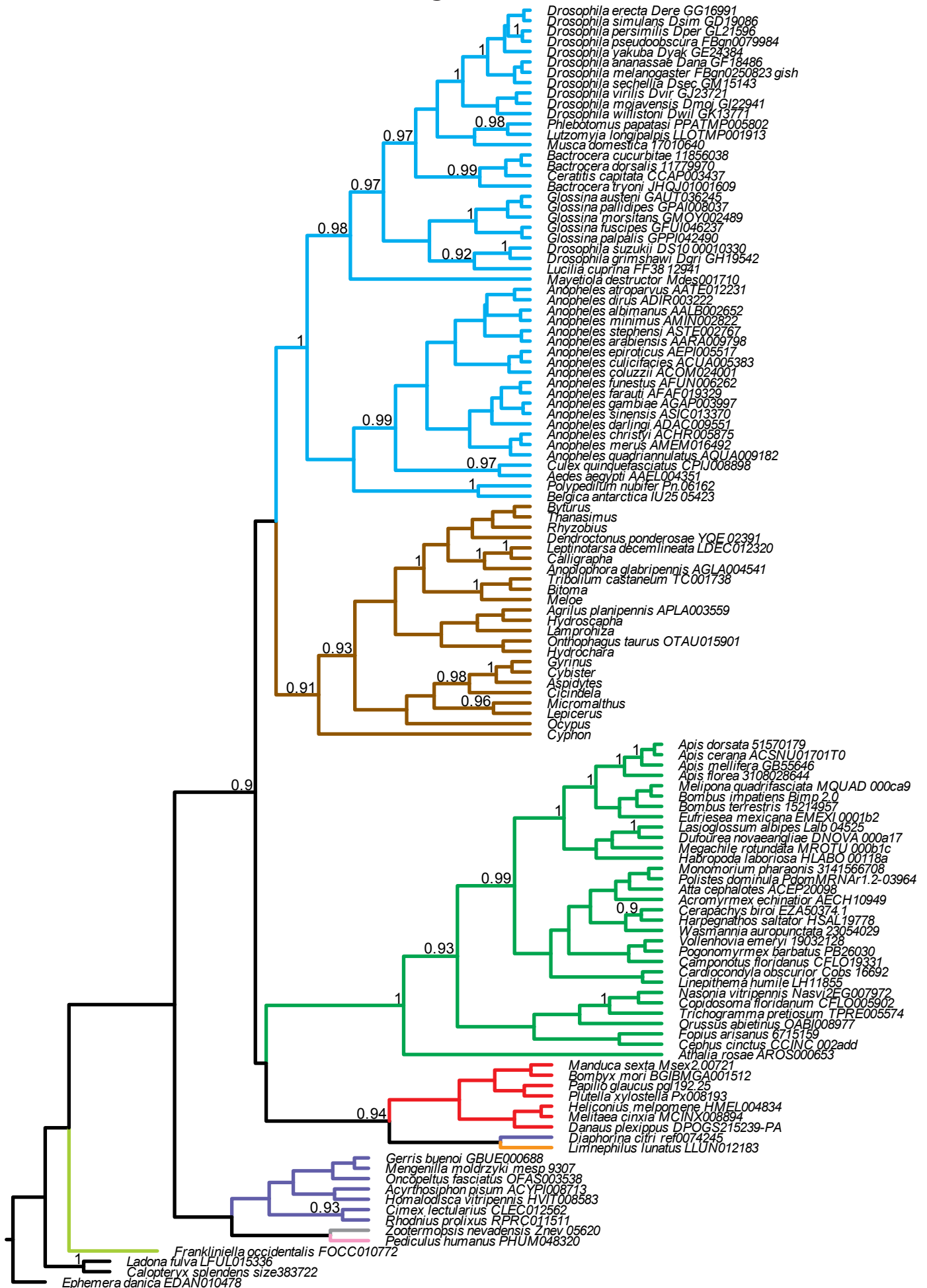
# eIF3m



# Fadd

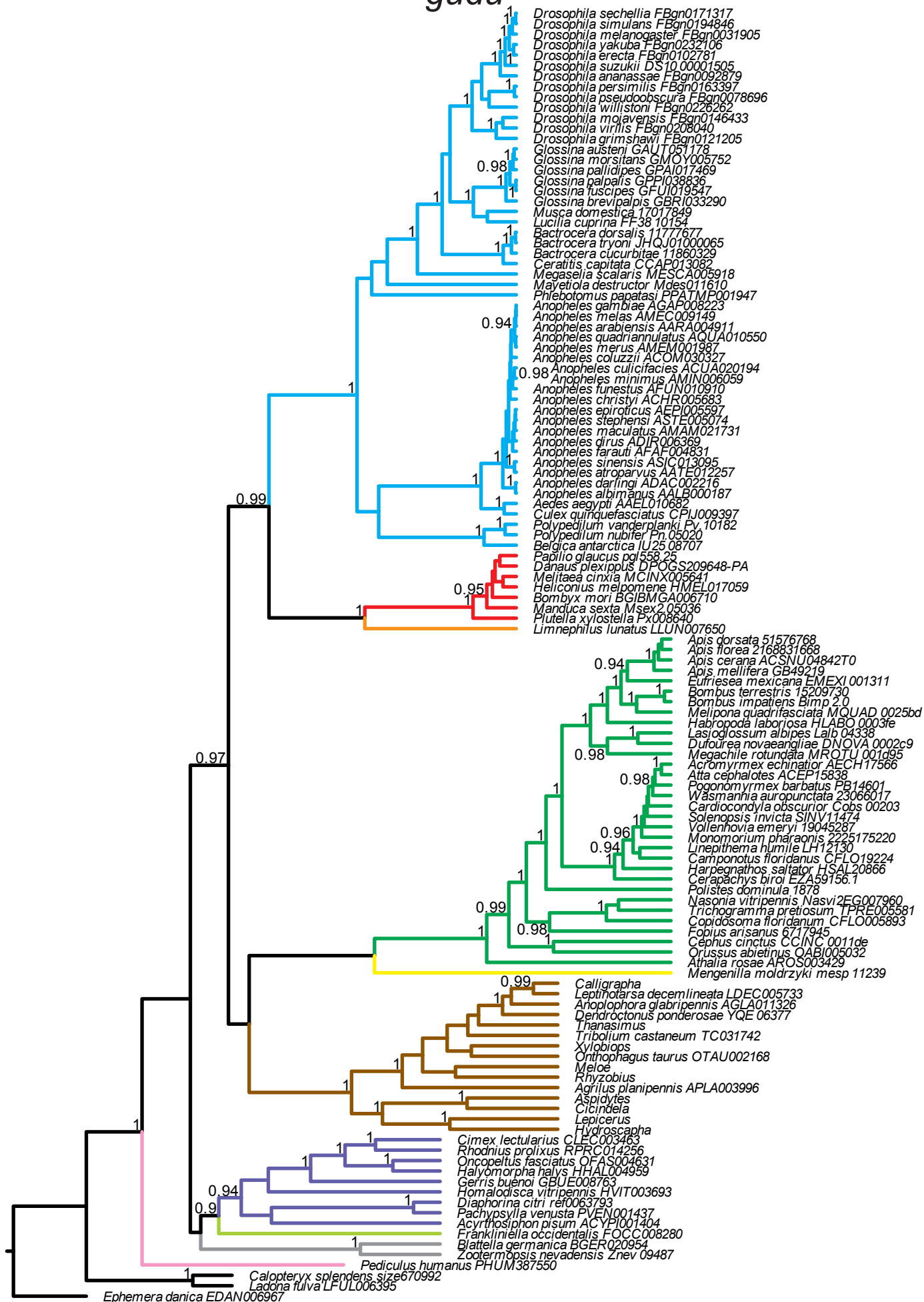


# gish

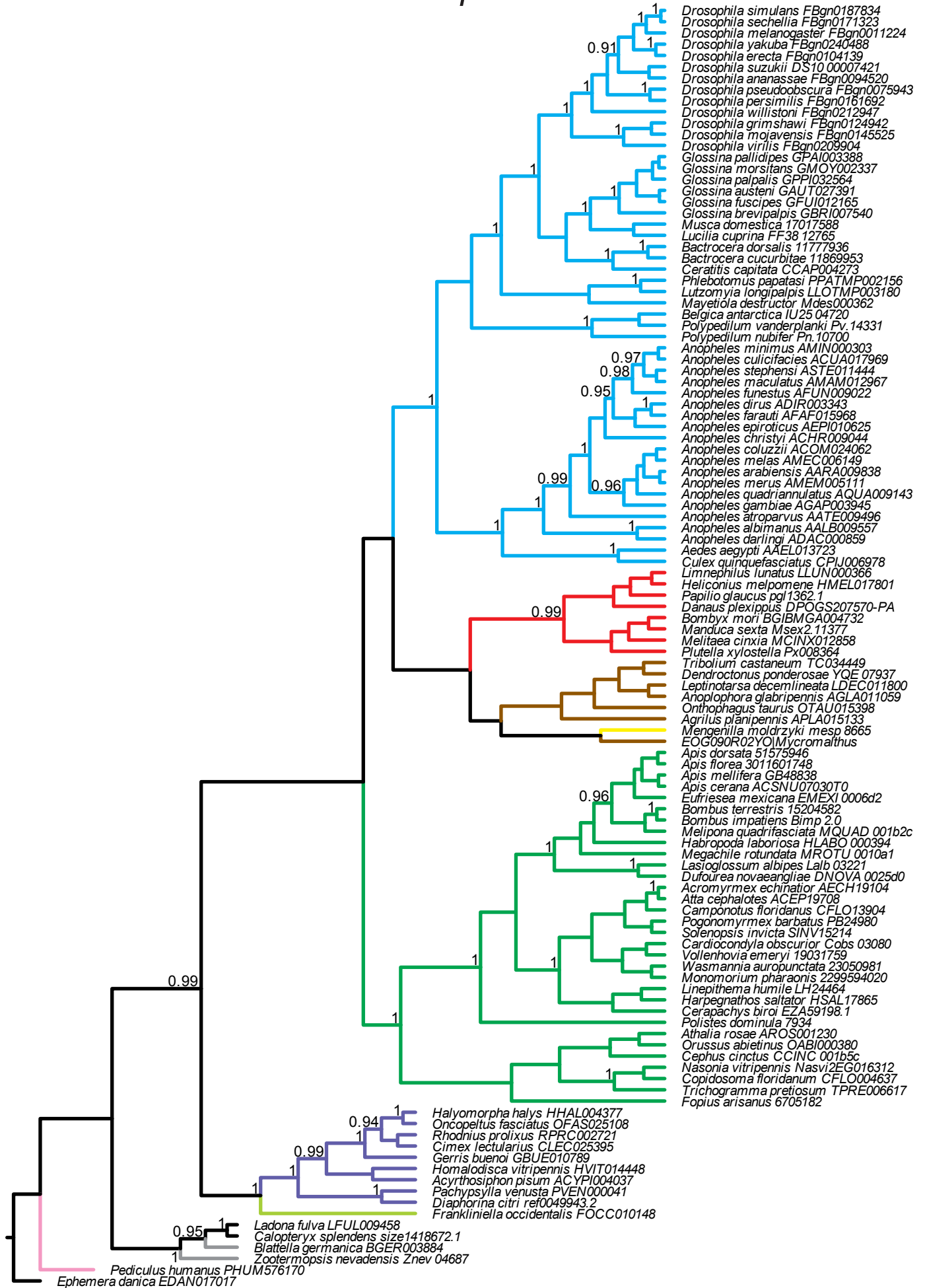




# gudu

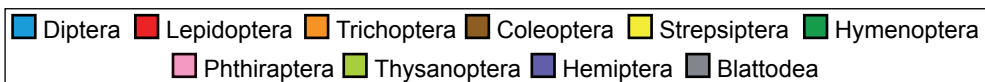
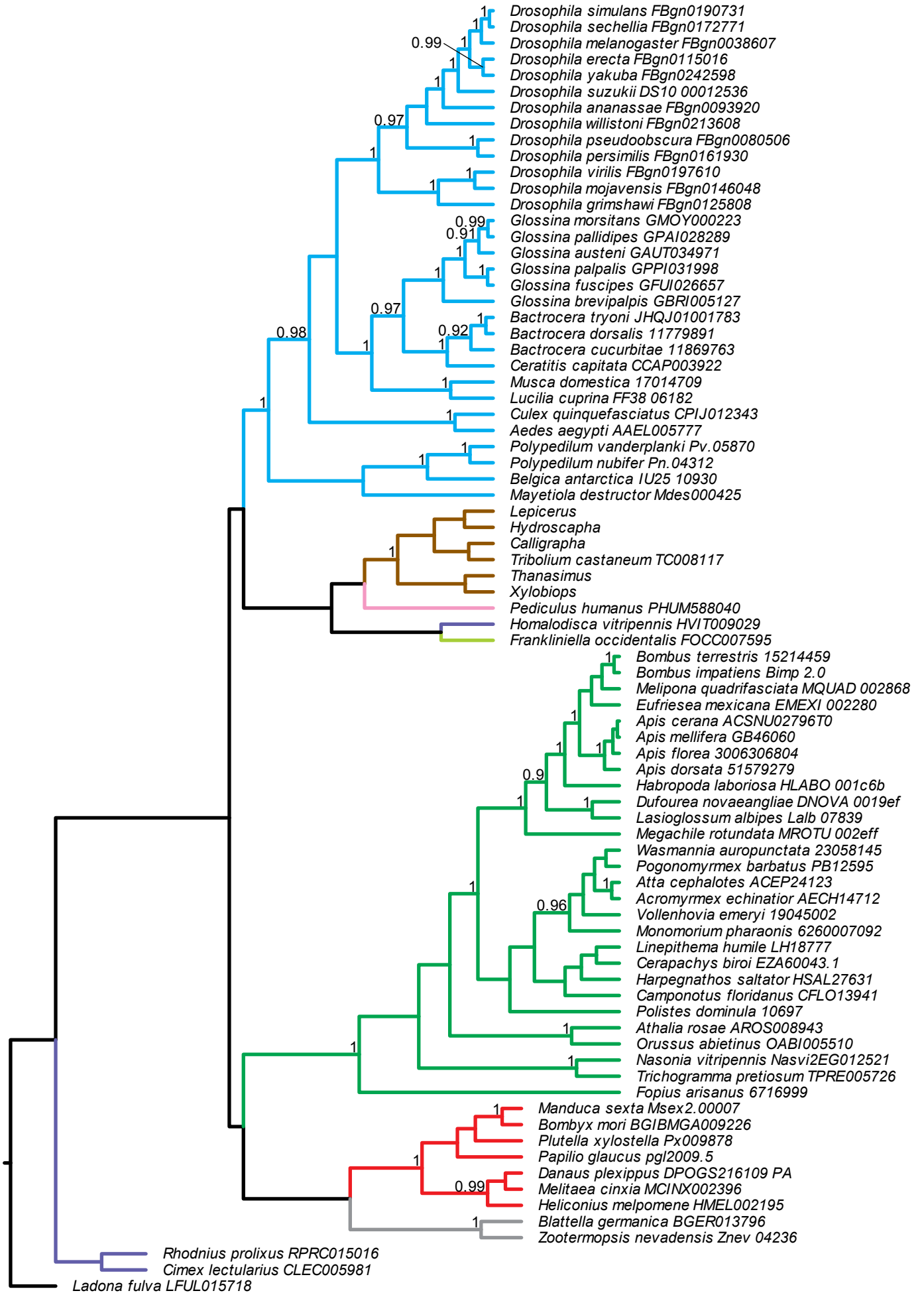


# heph

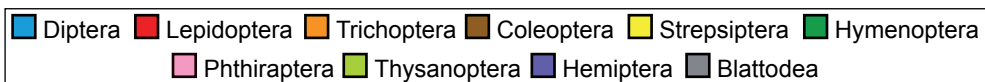
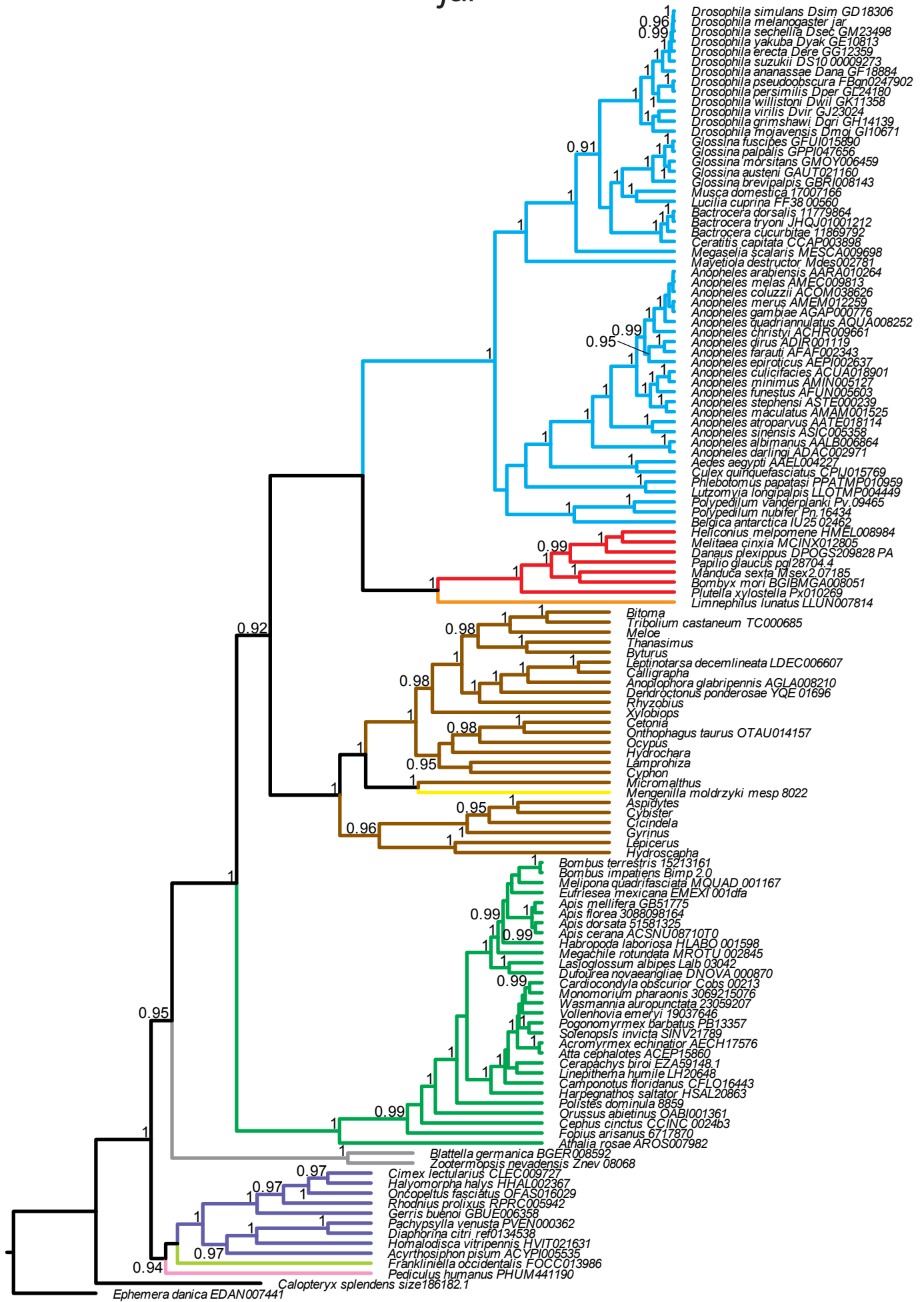




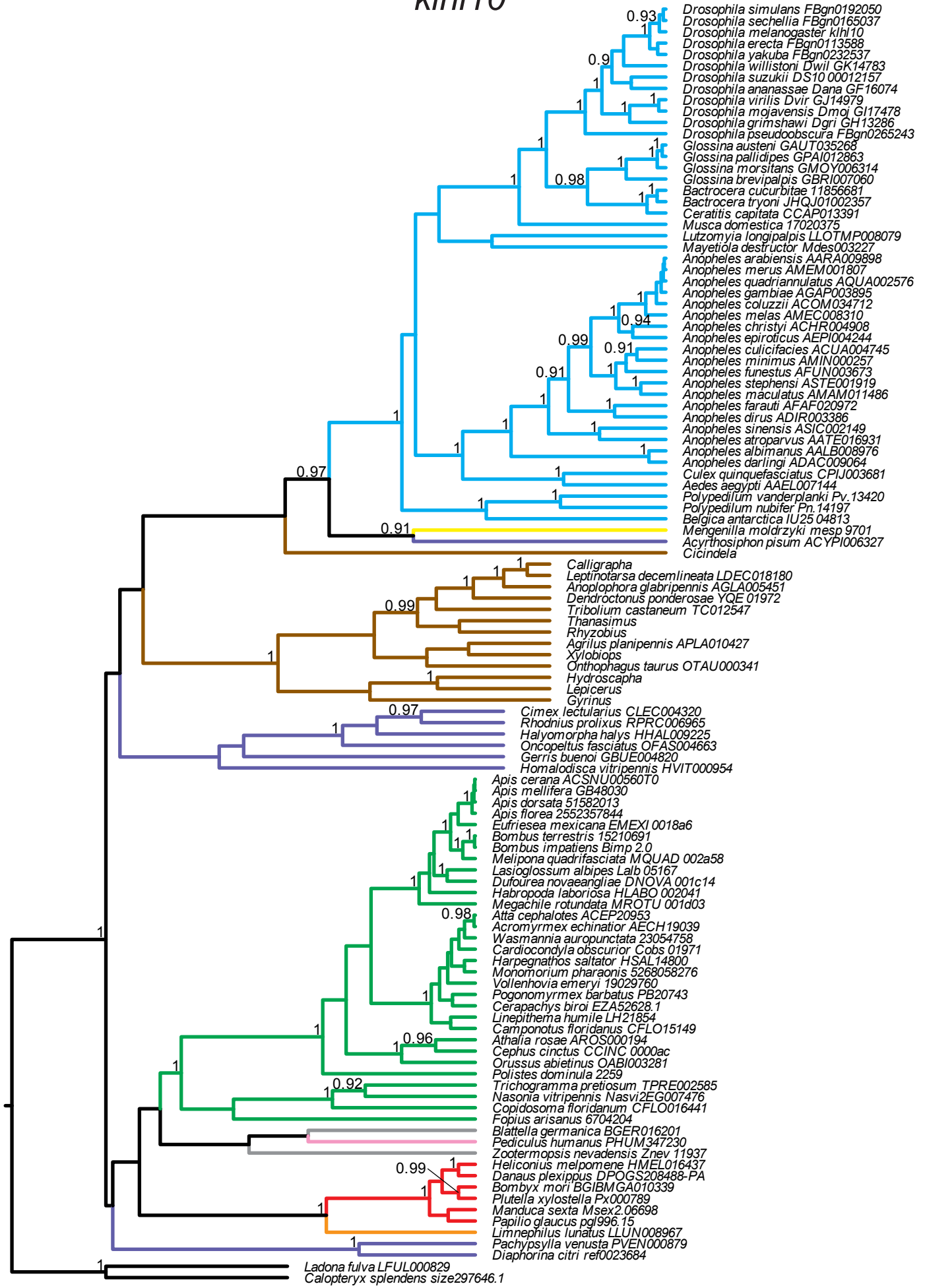
# hmw



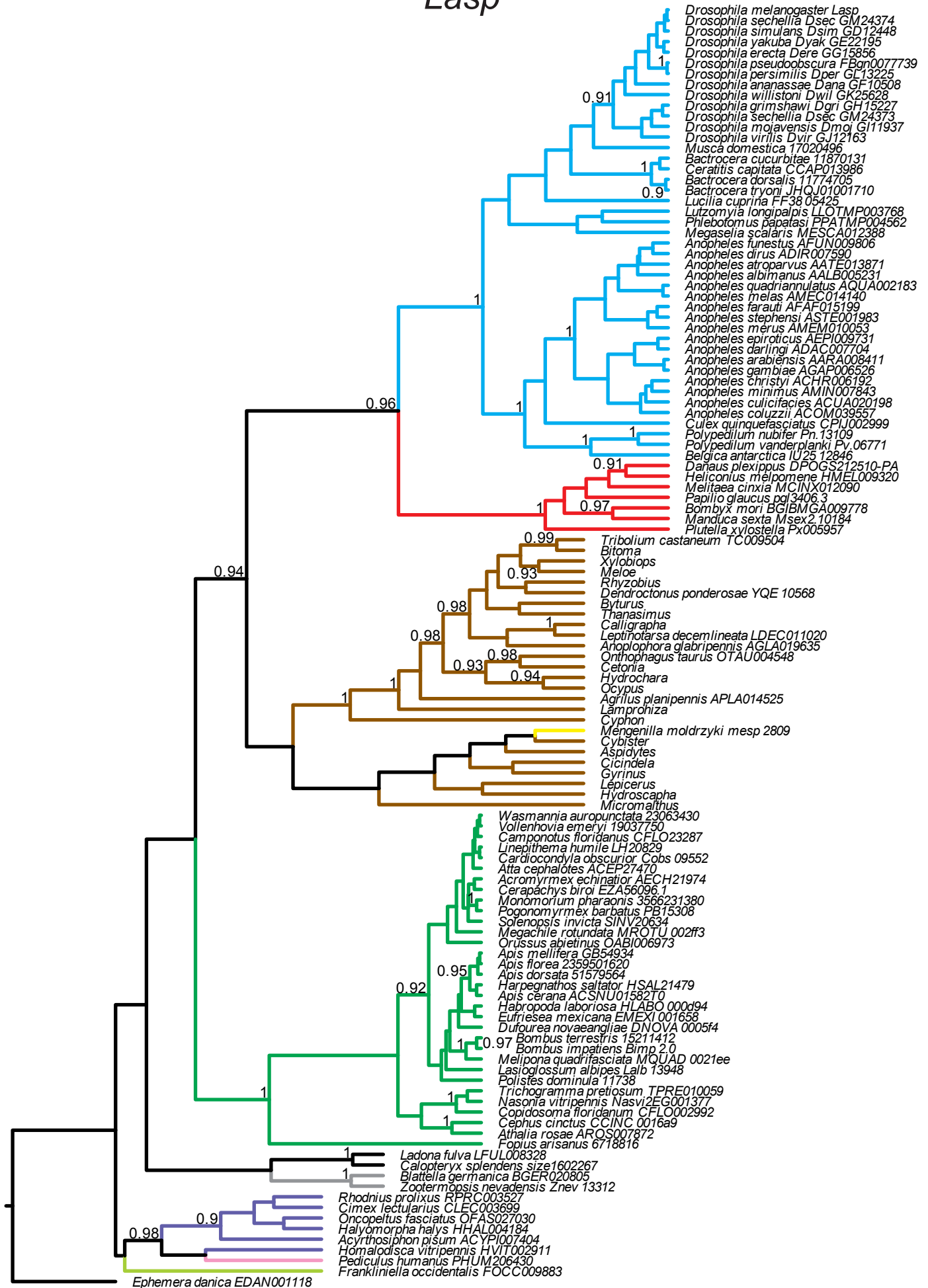
jar



# klhl10

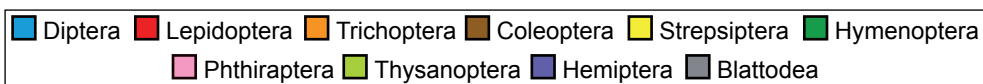
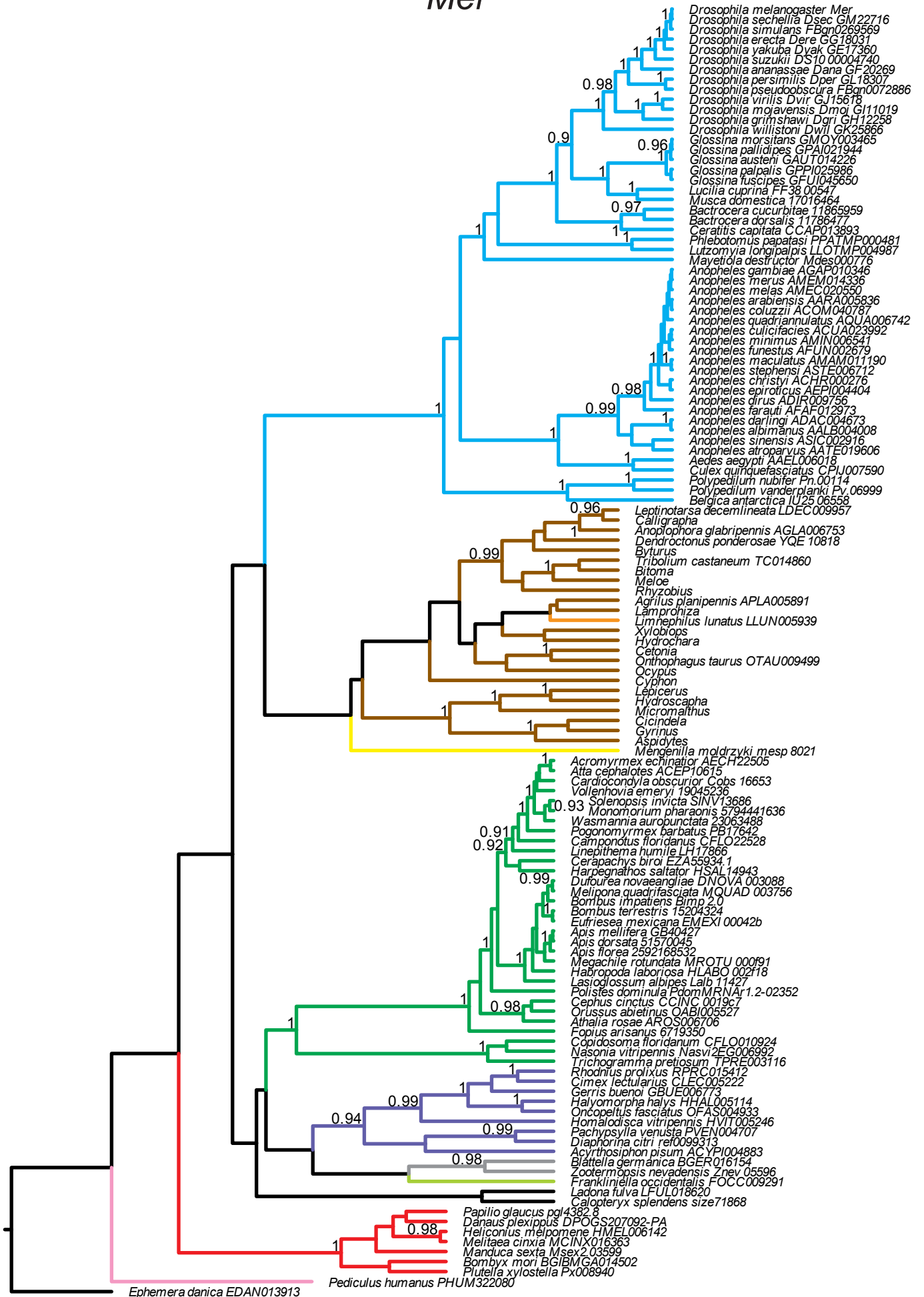


# Lasp

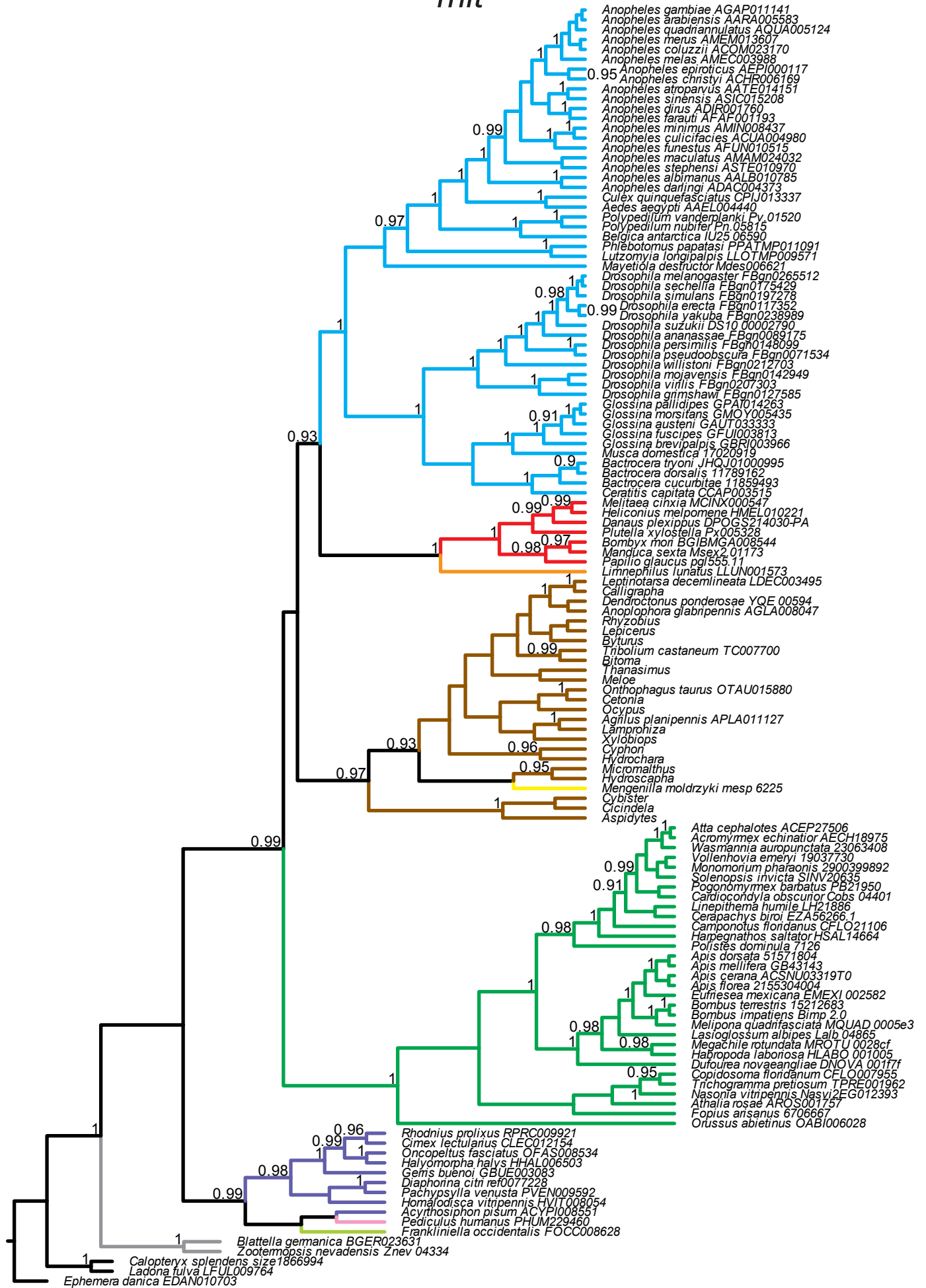




# Mer

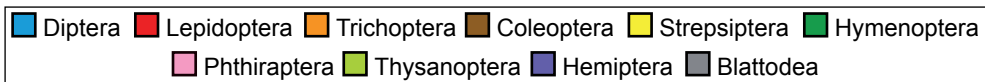
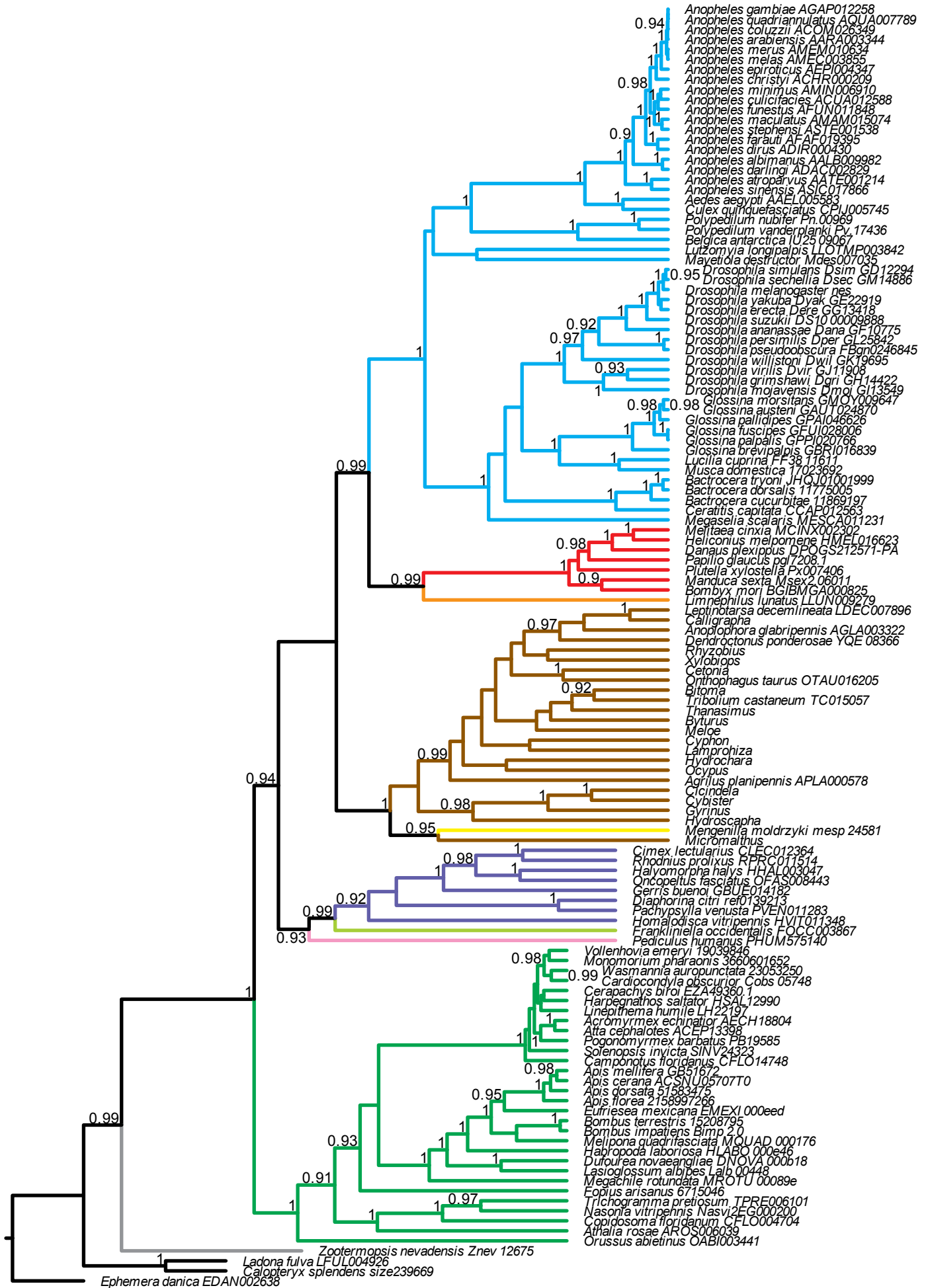


mlt

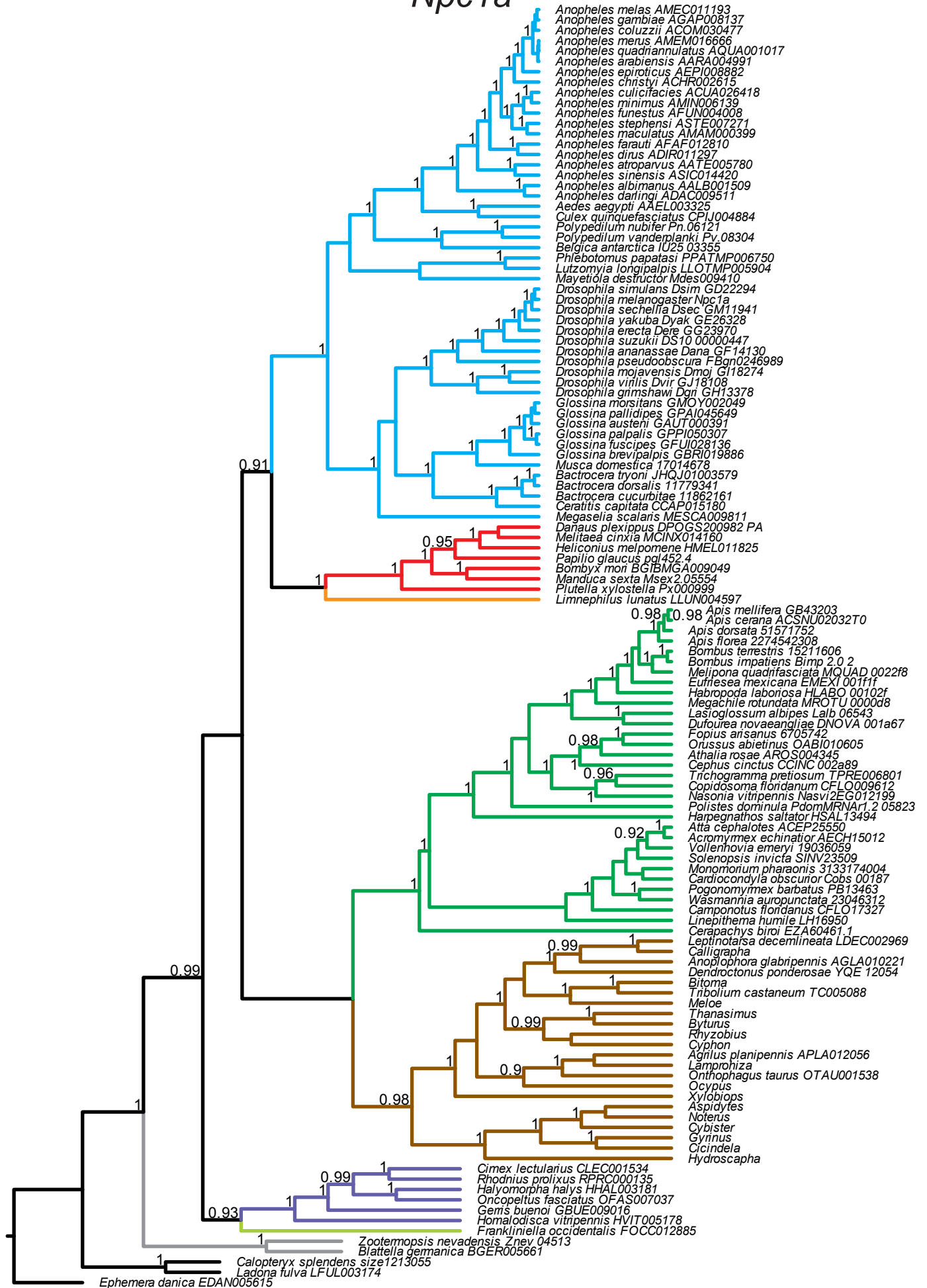




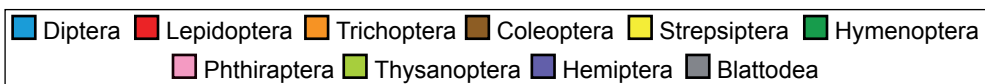
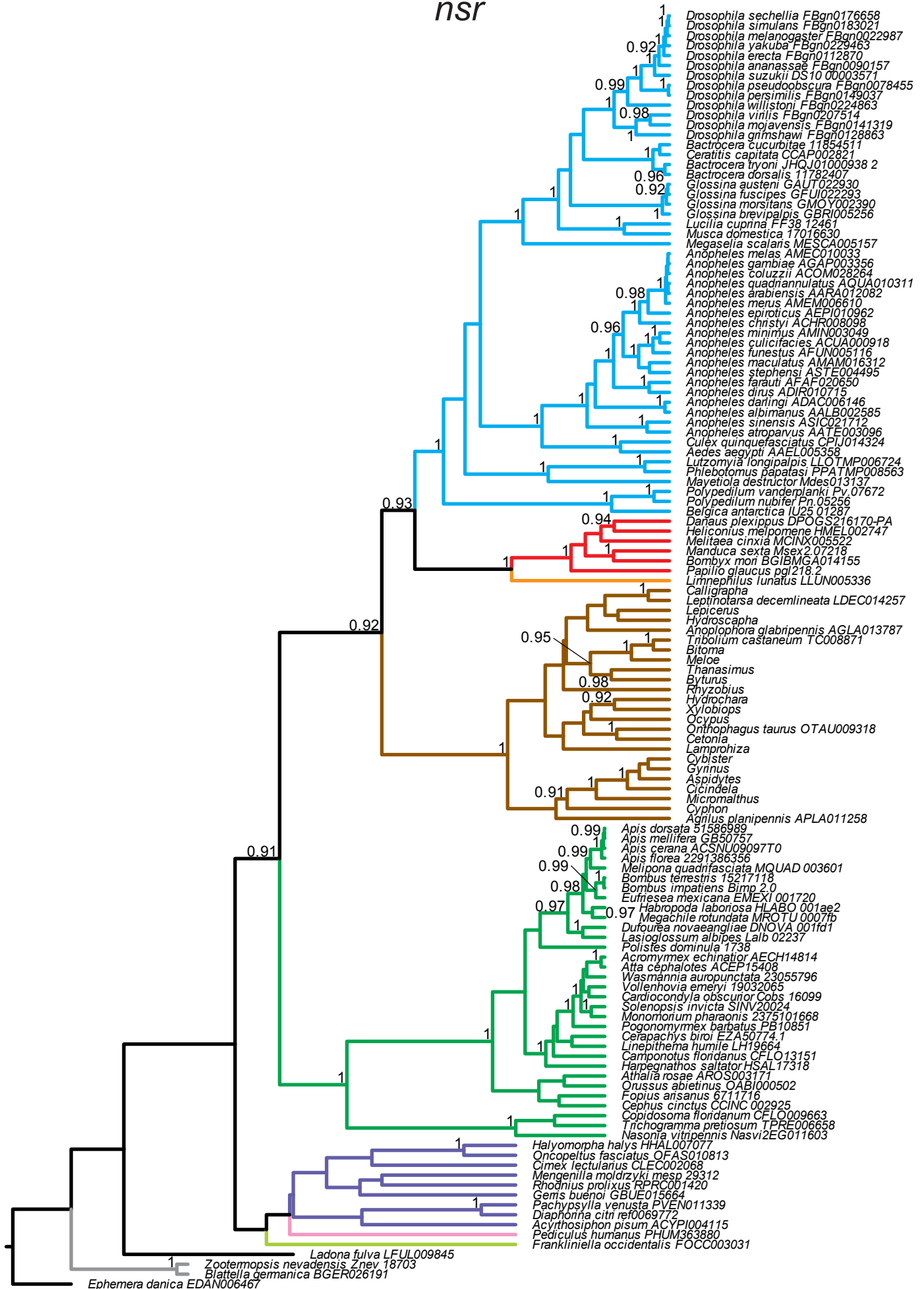
nes



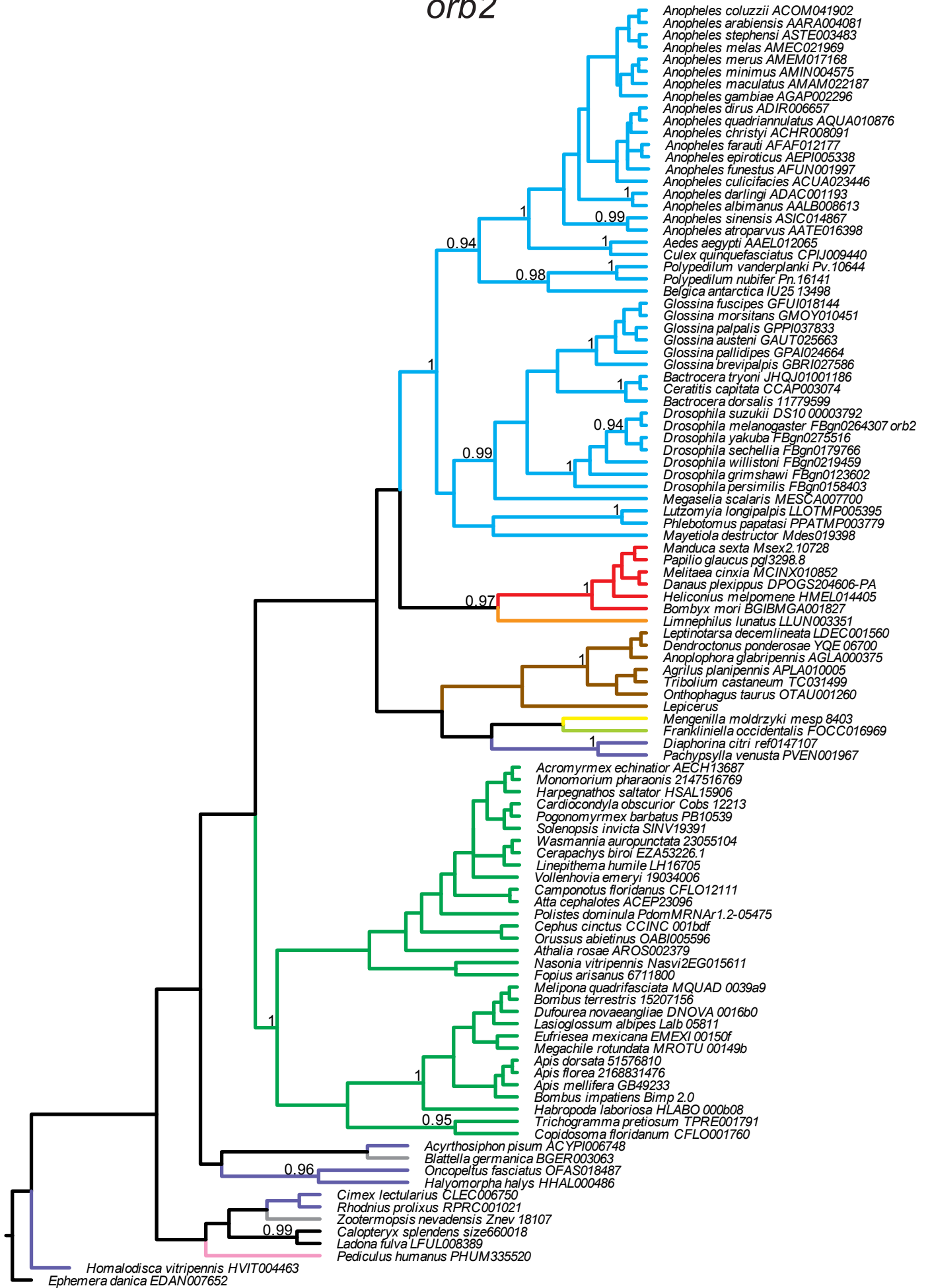
# Npc1a



nsr

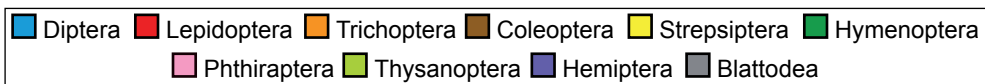
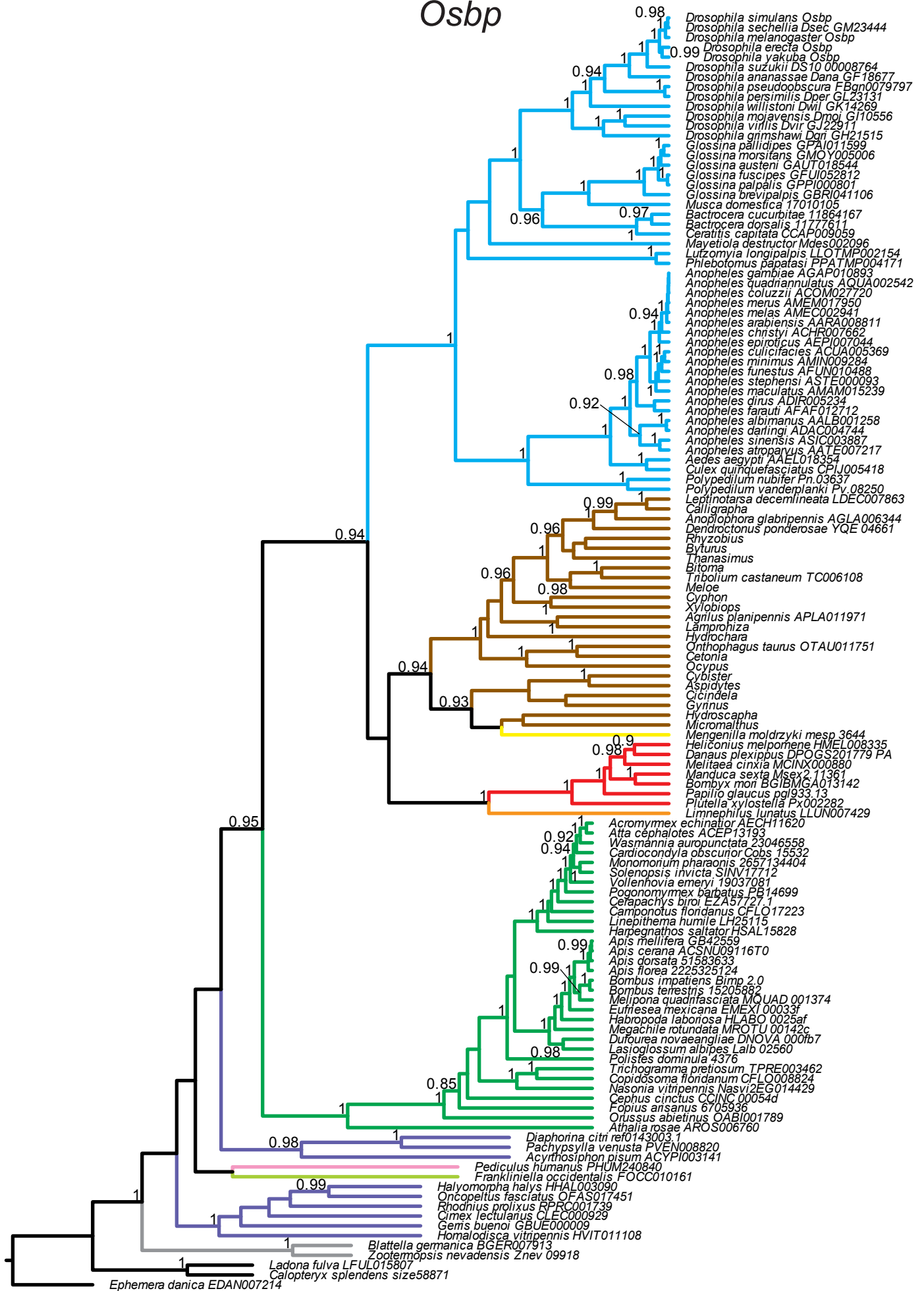


# orb2

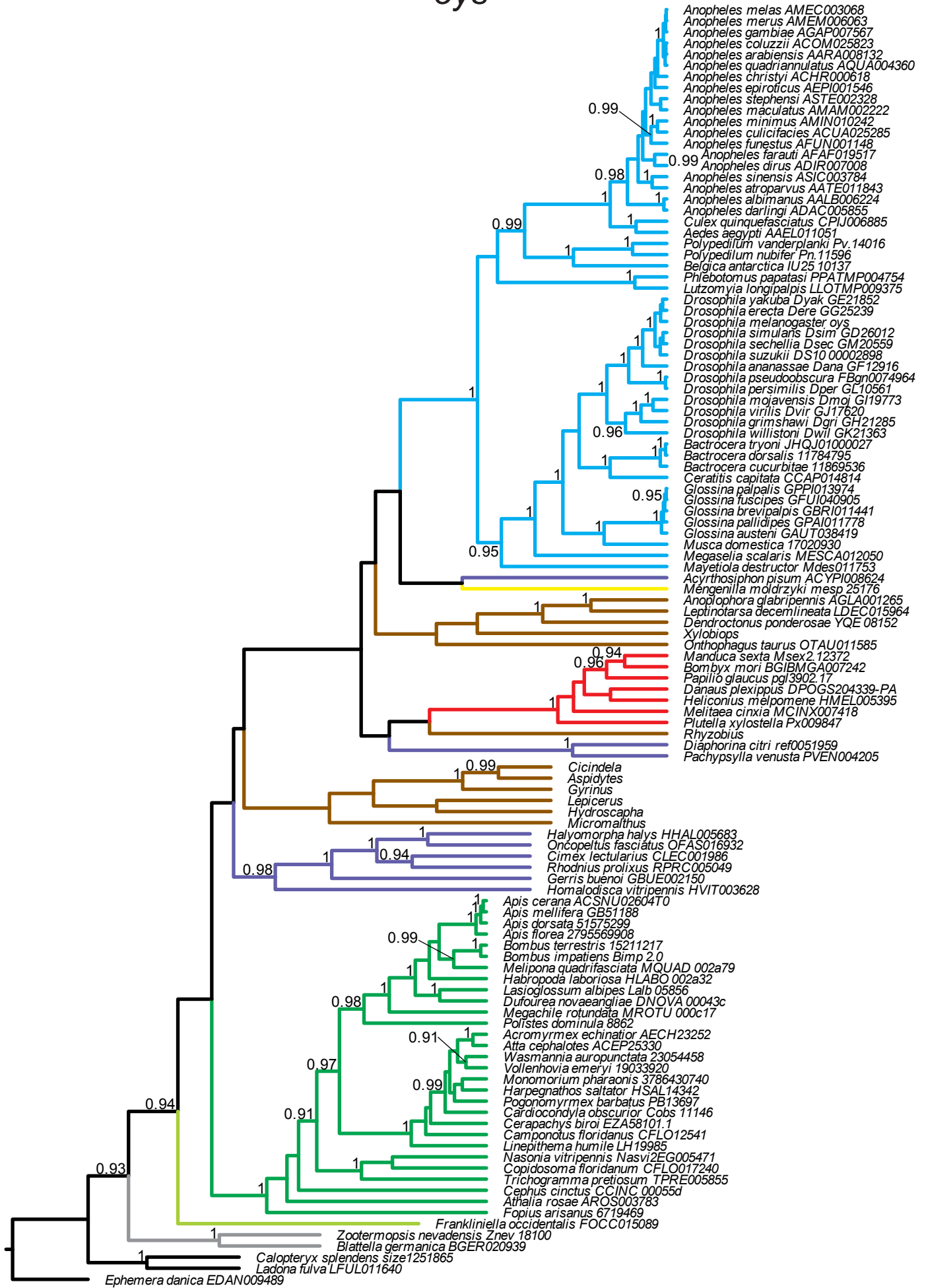




# Osbp

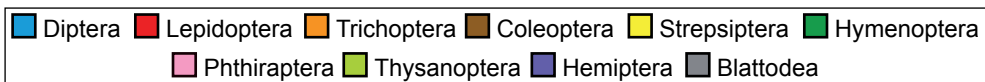
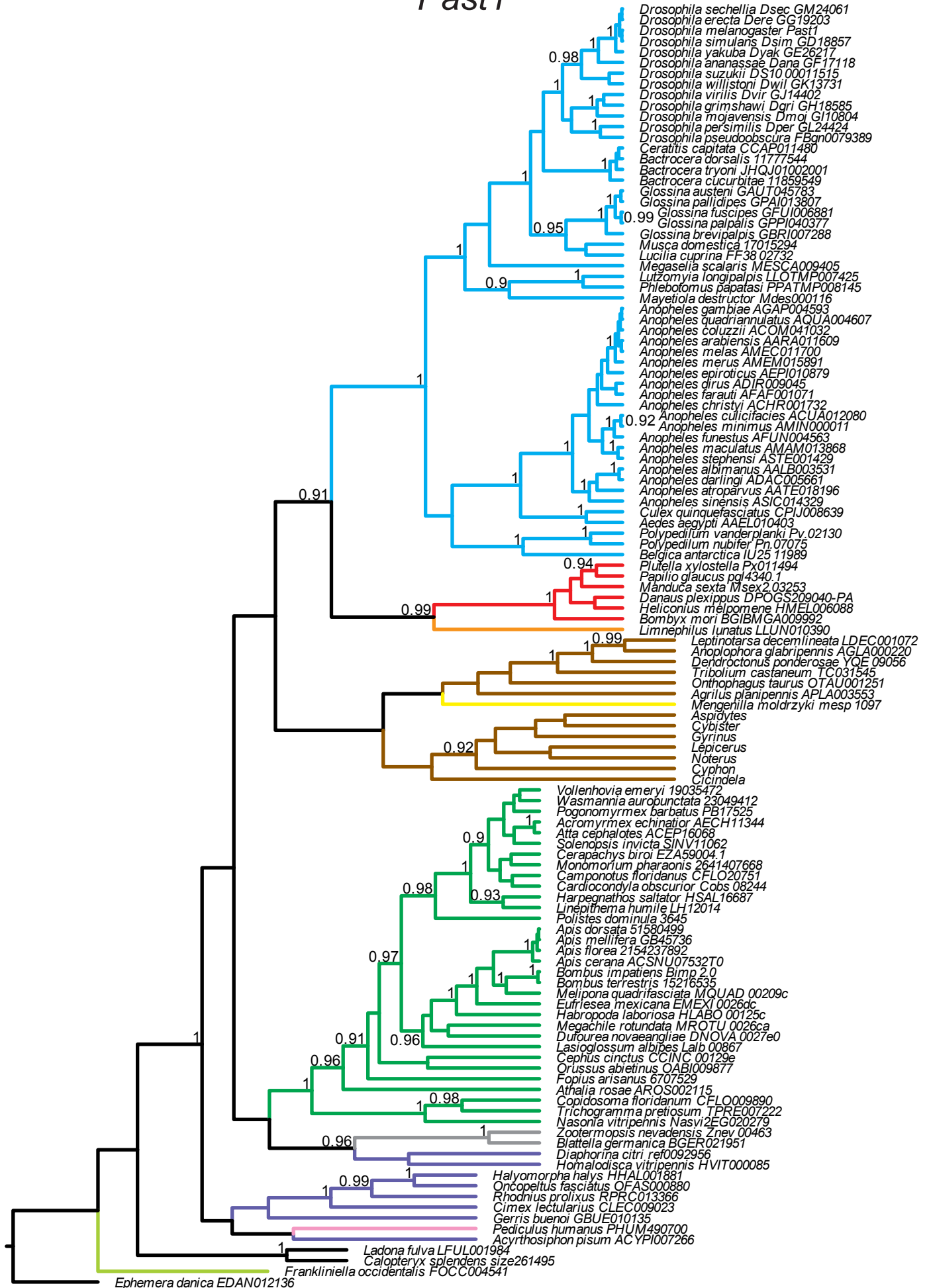


# oys

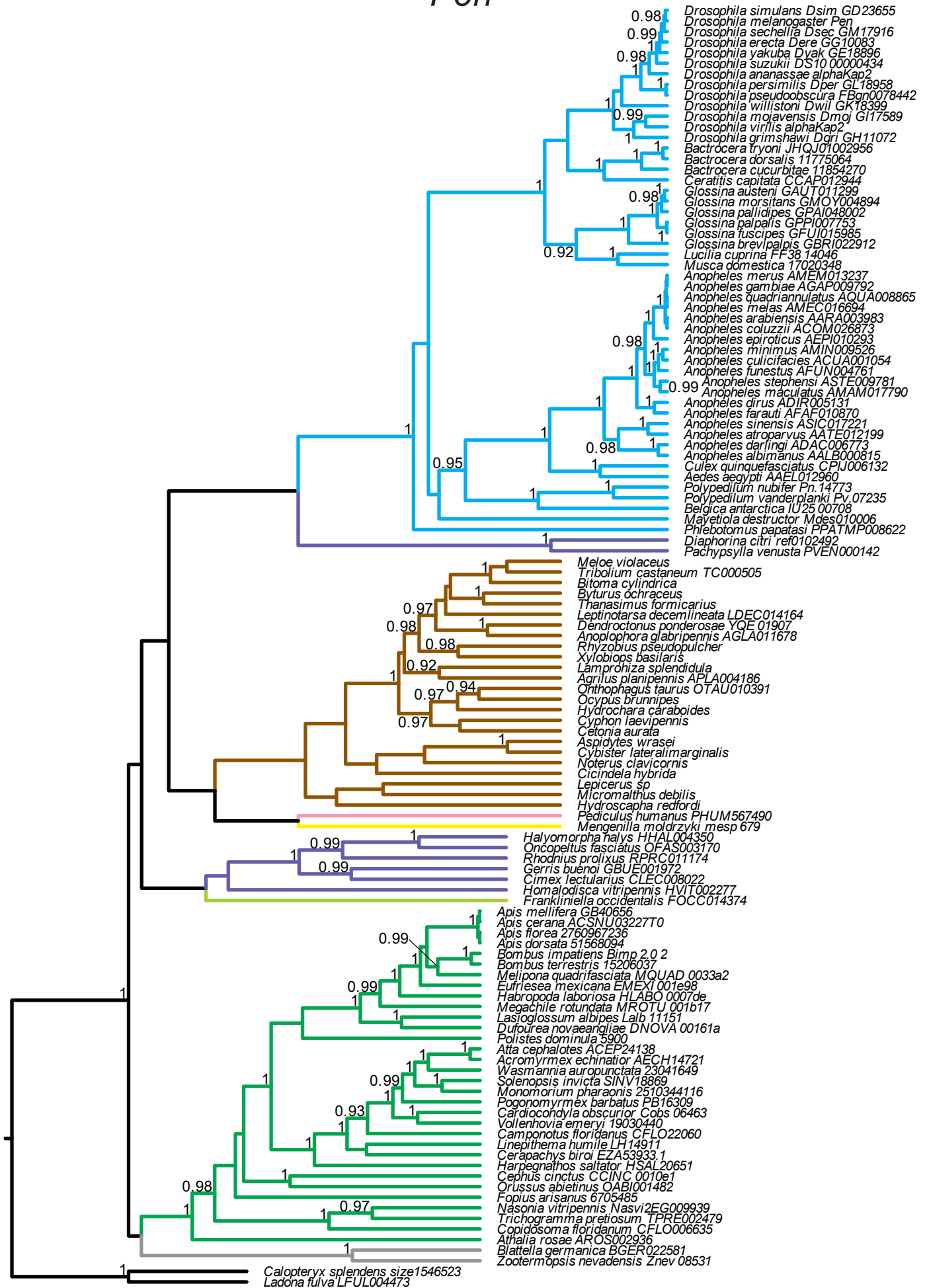




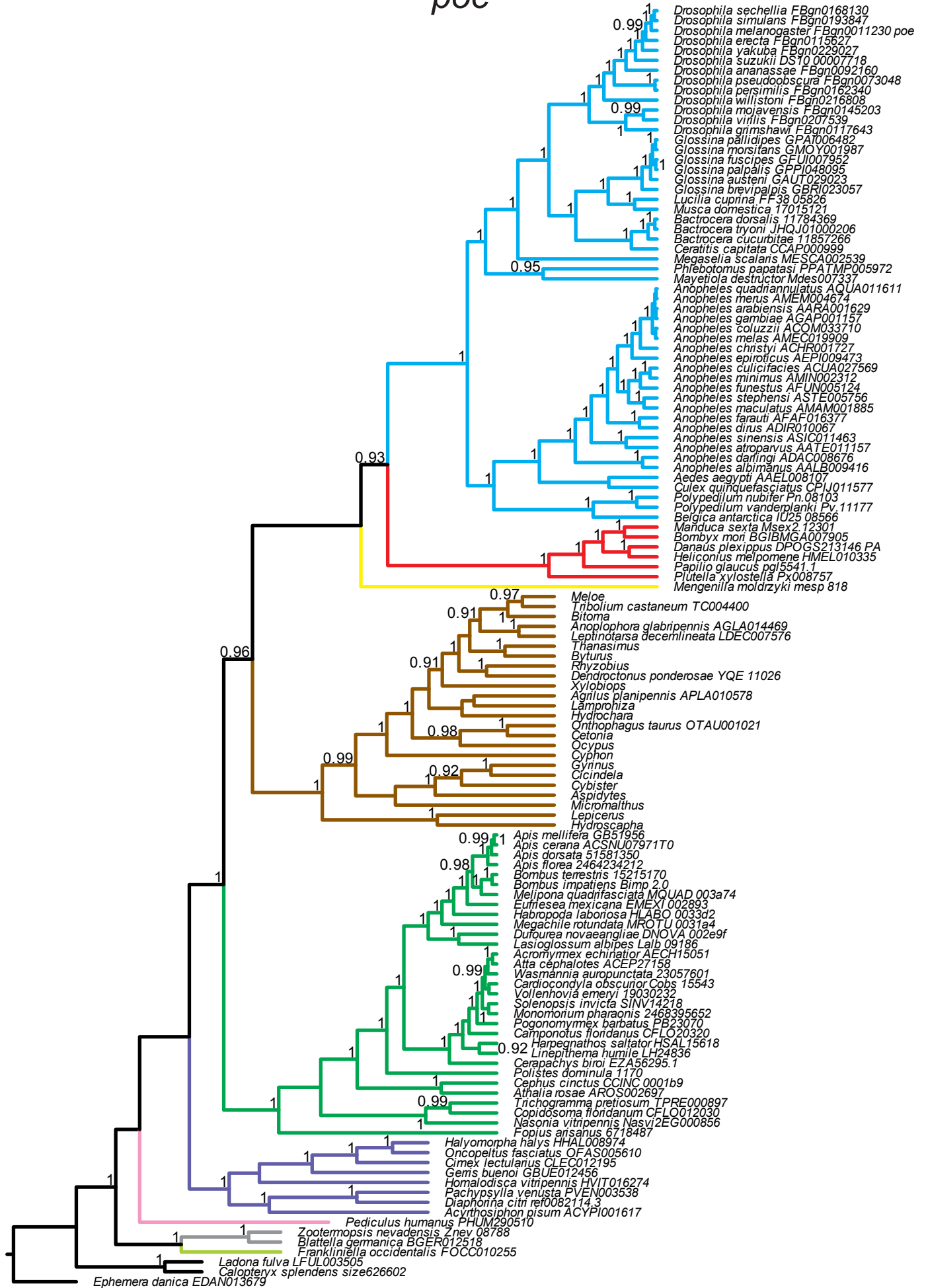
# Past1



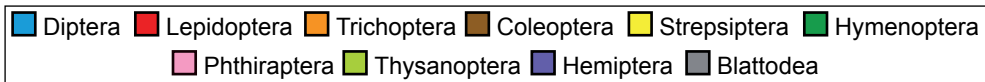
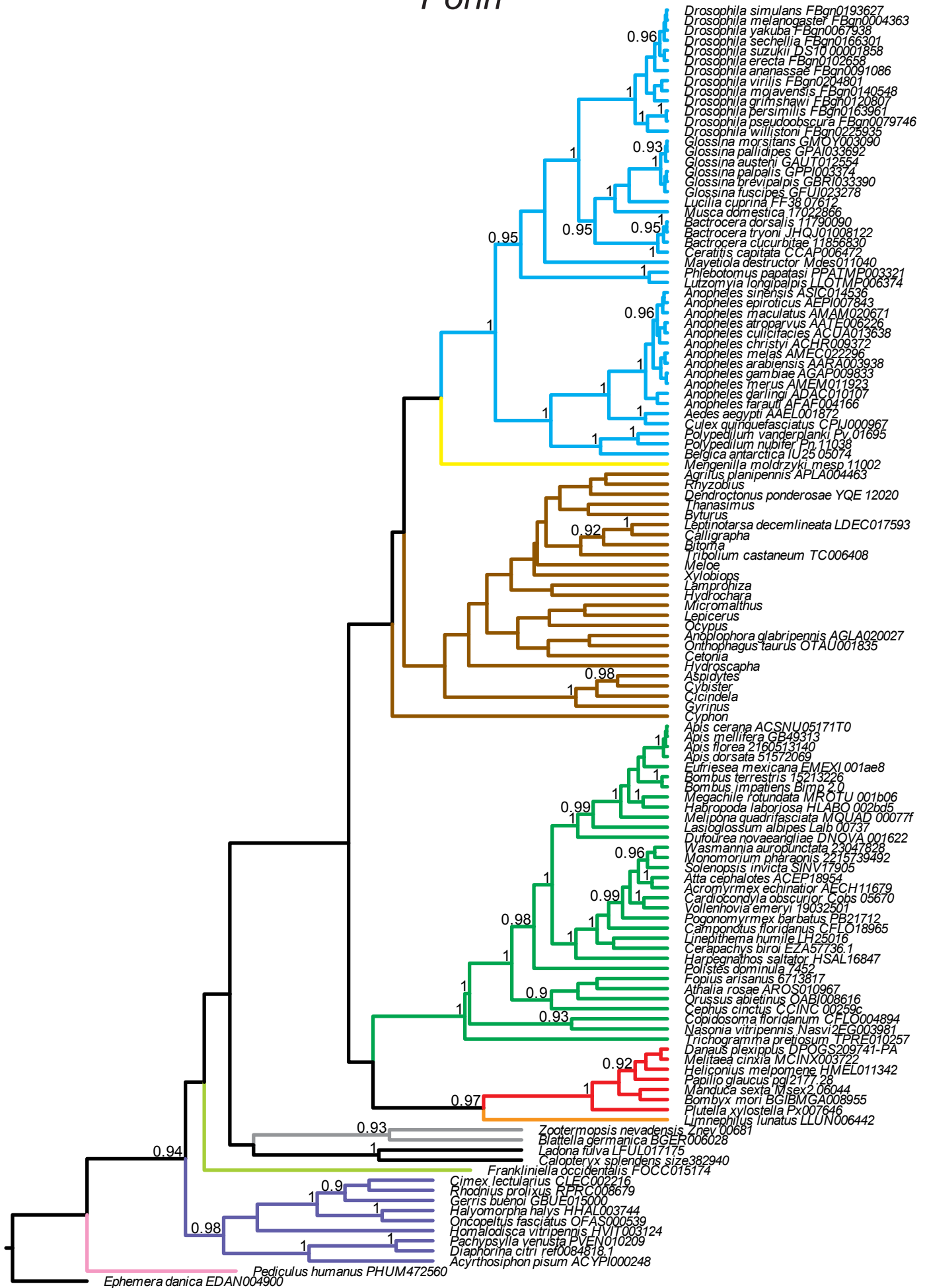
# Pen



poe

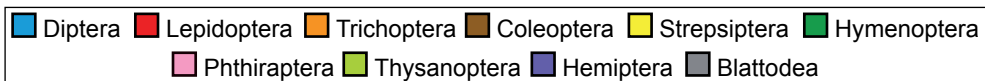
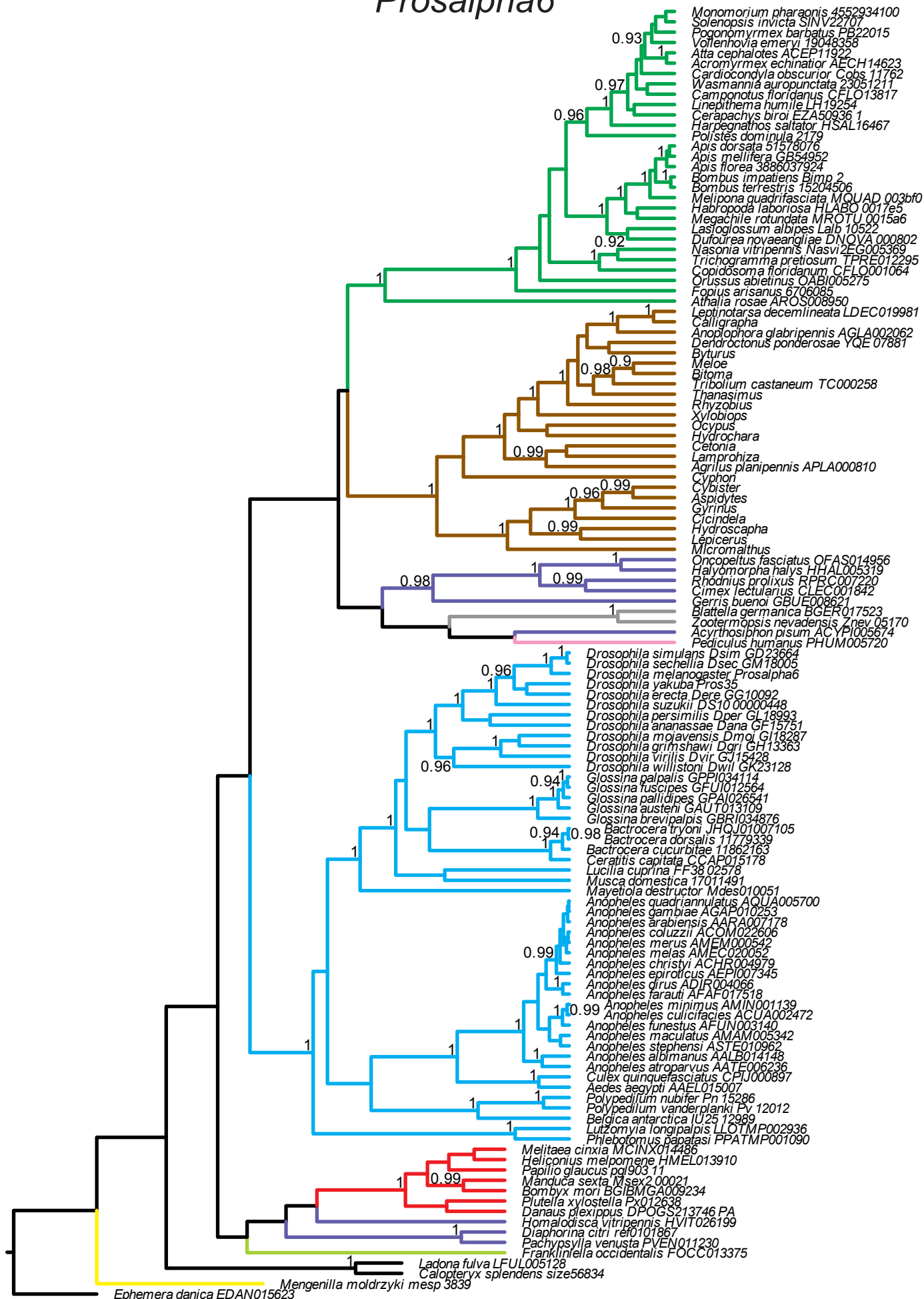


# Porin

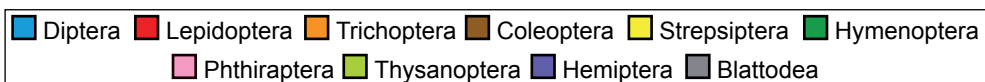
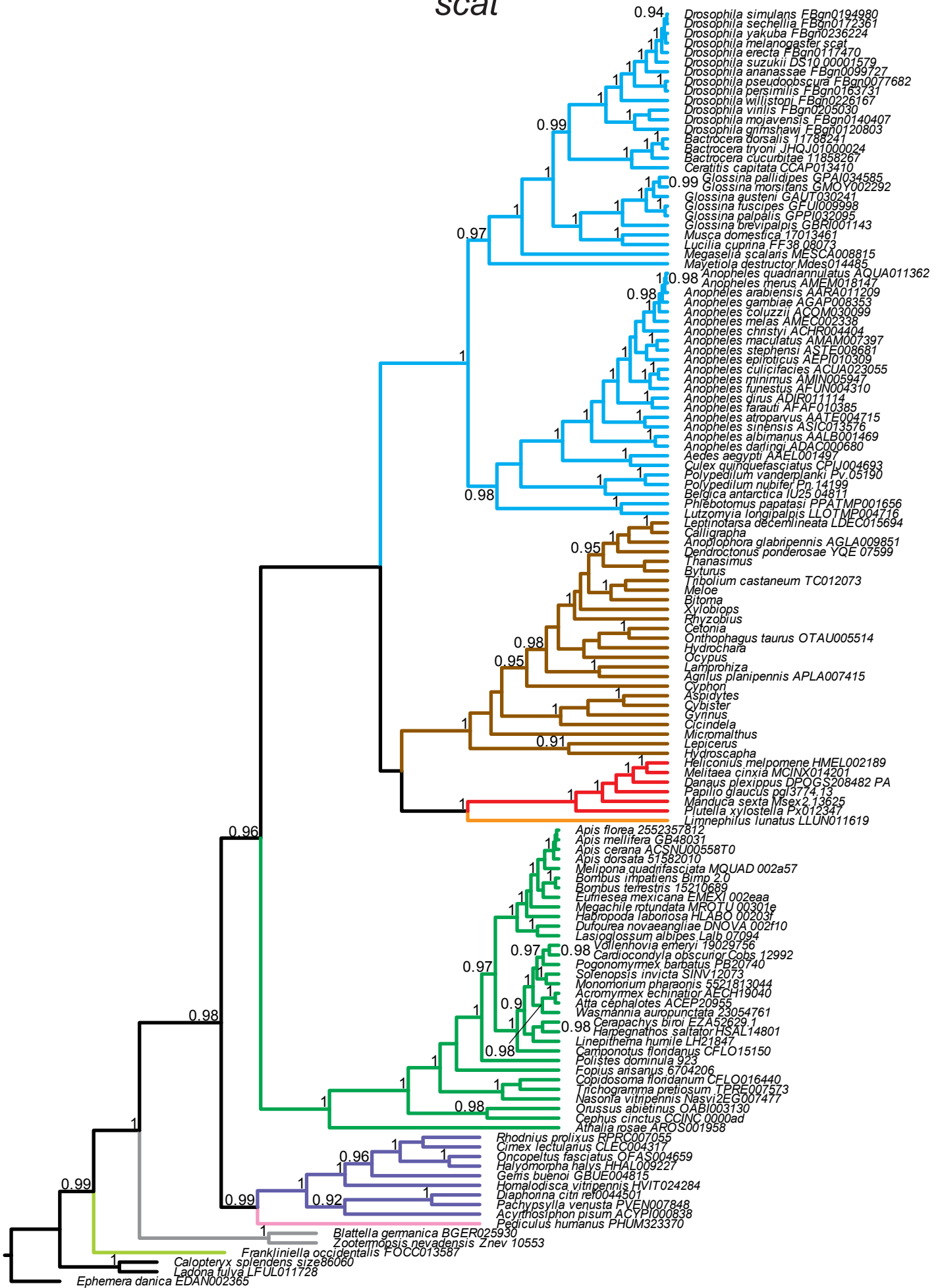




# Prosalpha6

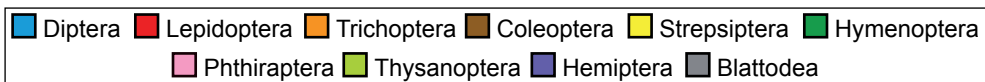
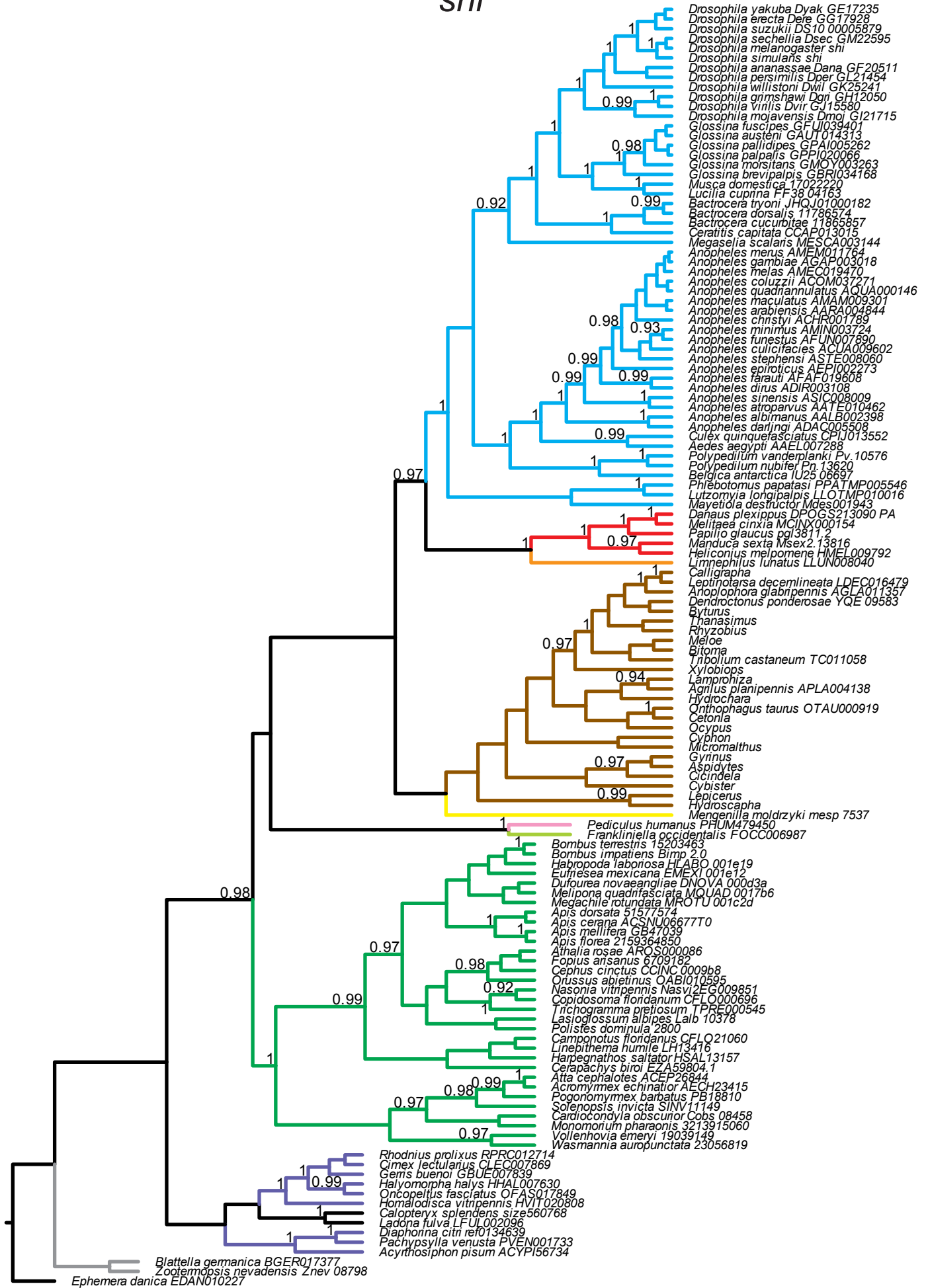


# scat

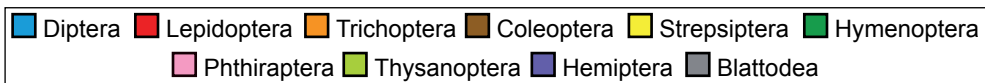
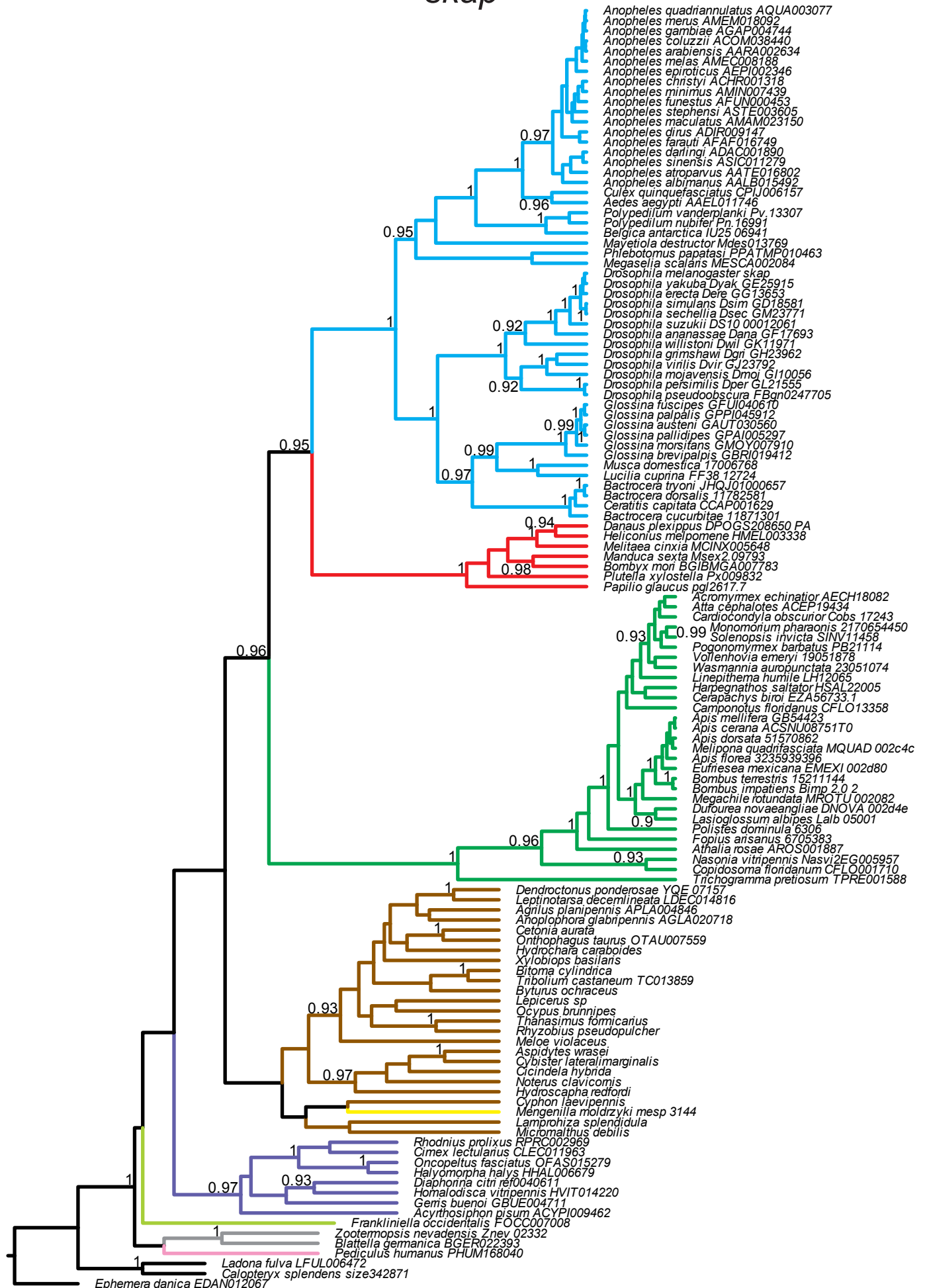




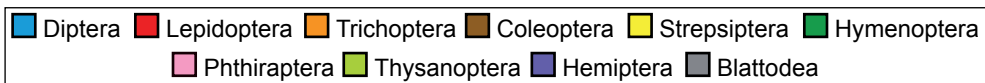
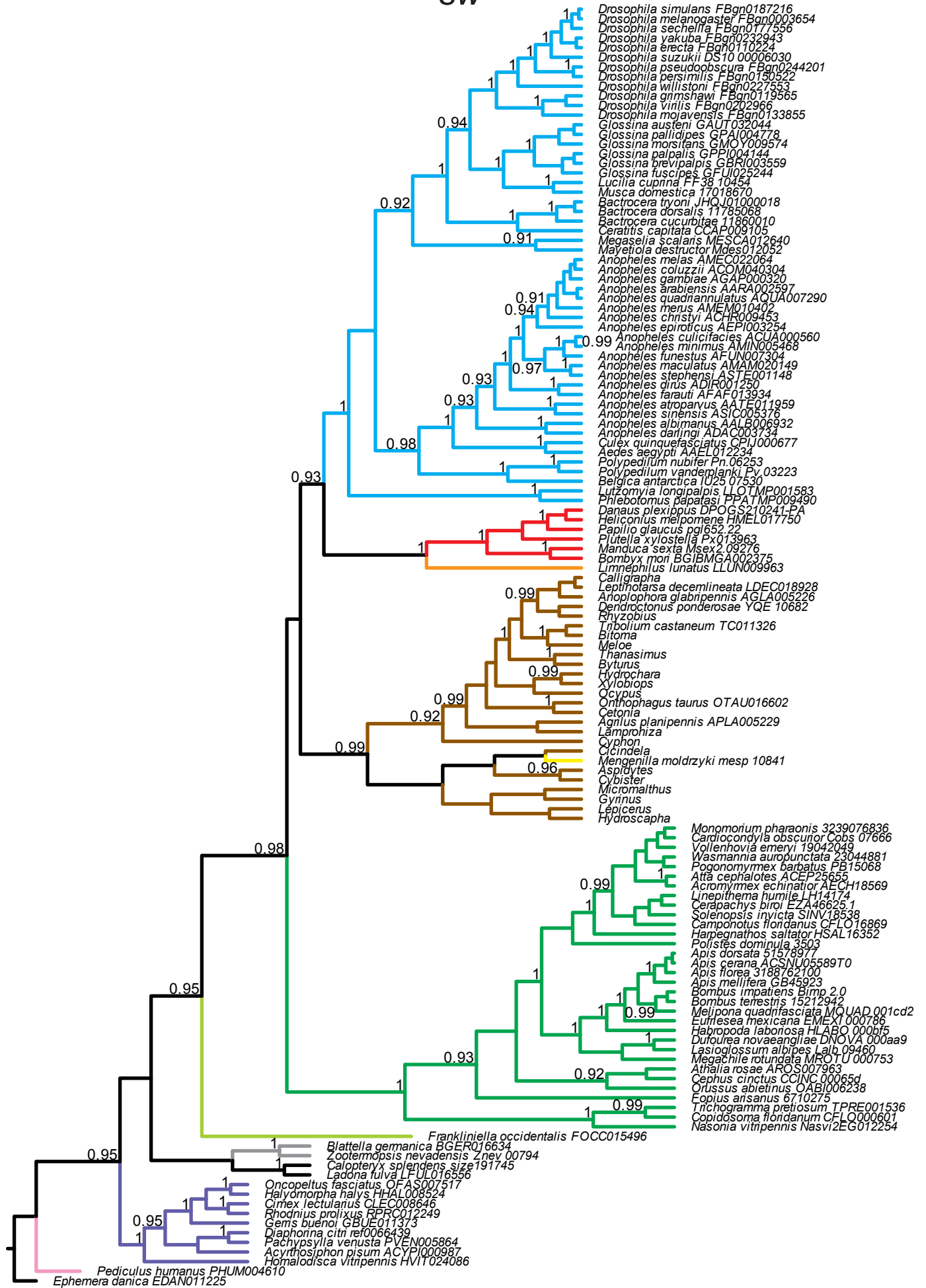
# shi



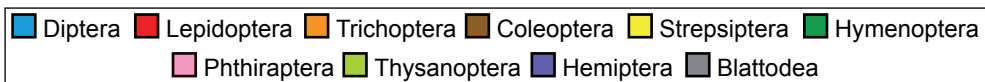
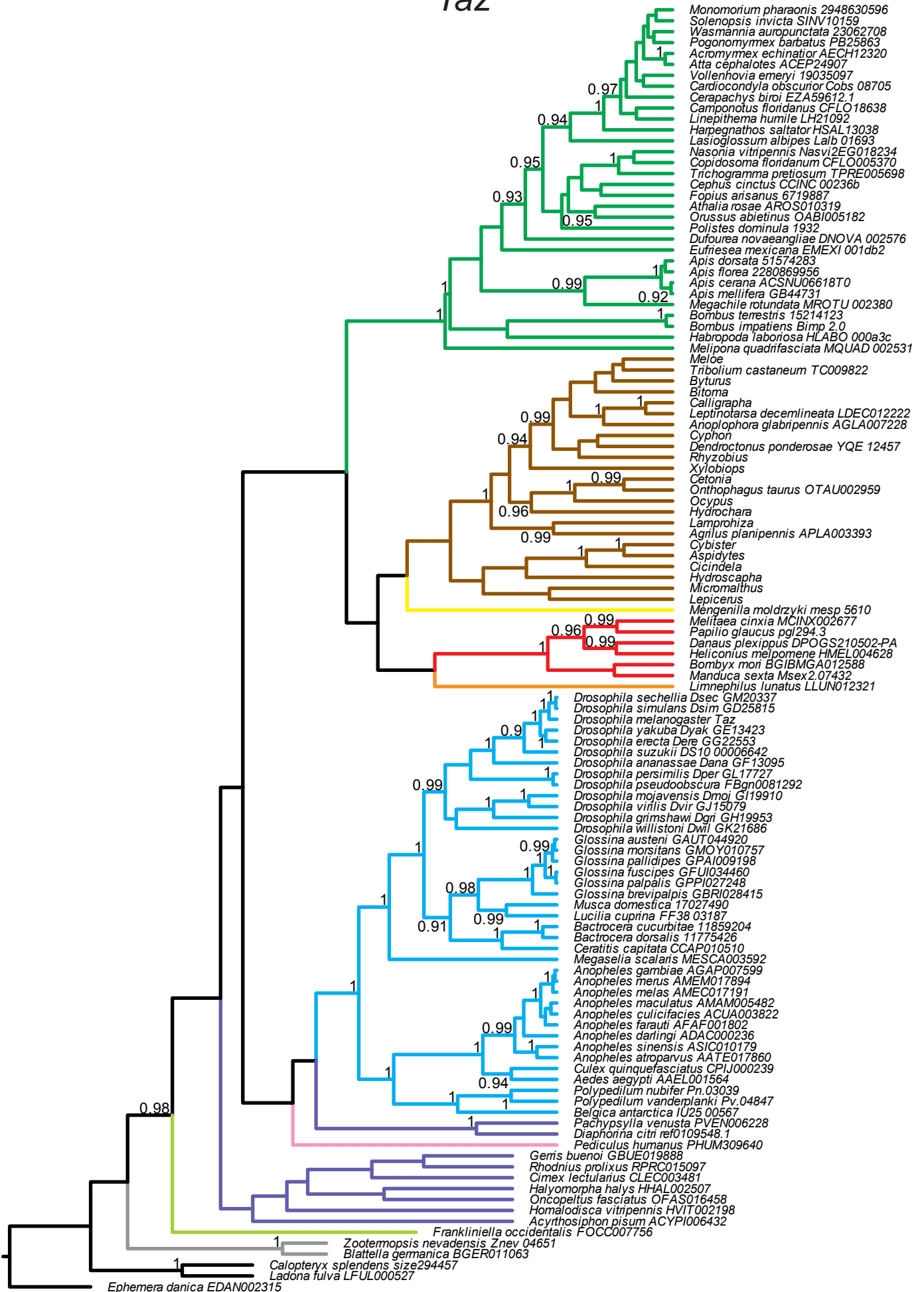
# skap



SW

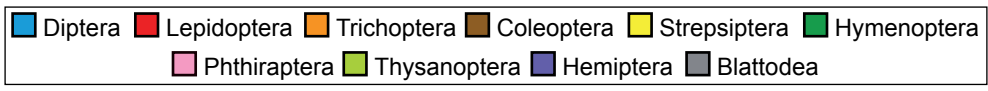
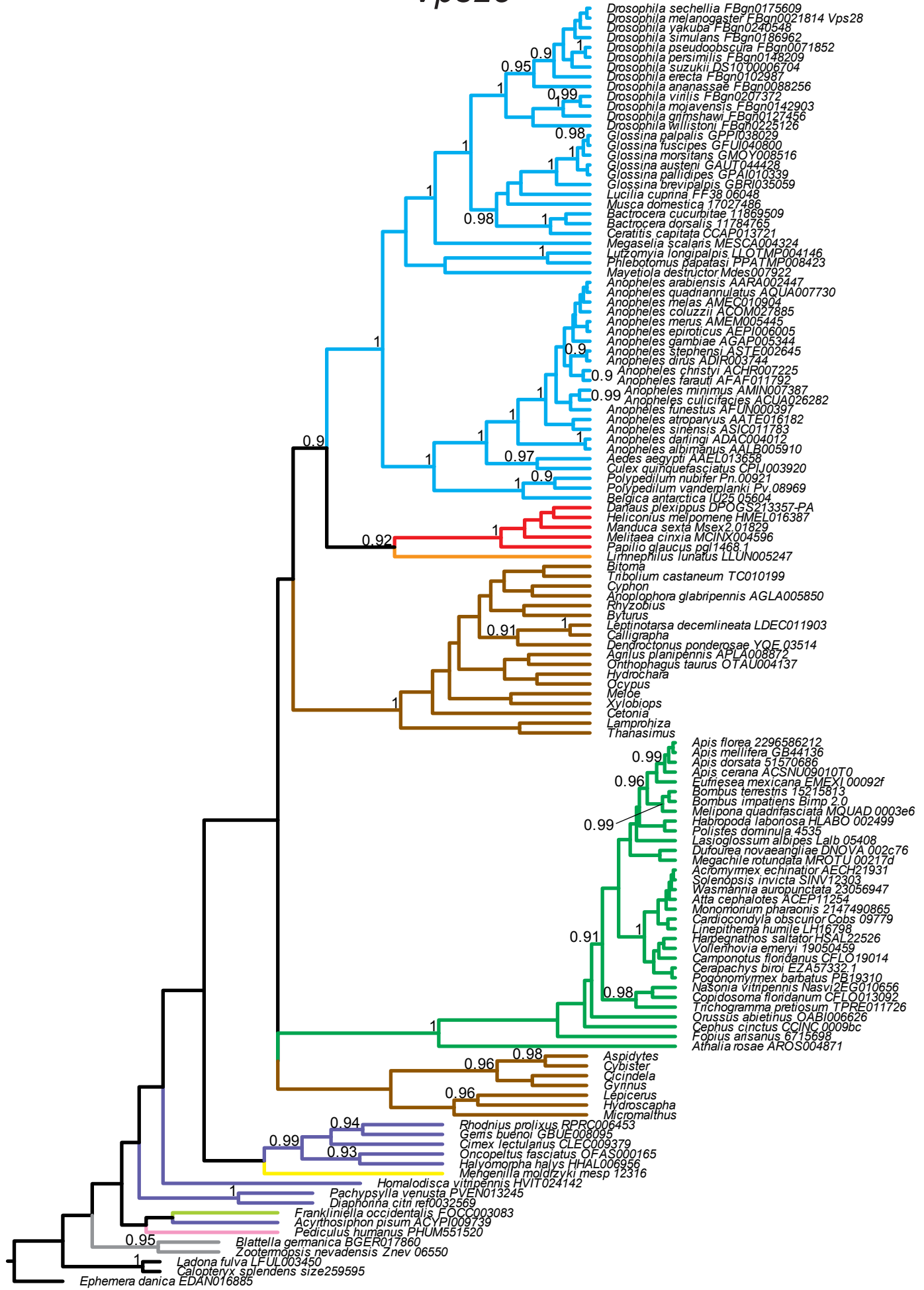


# Taz





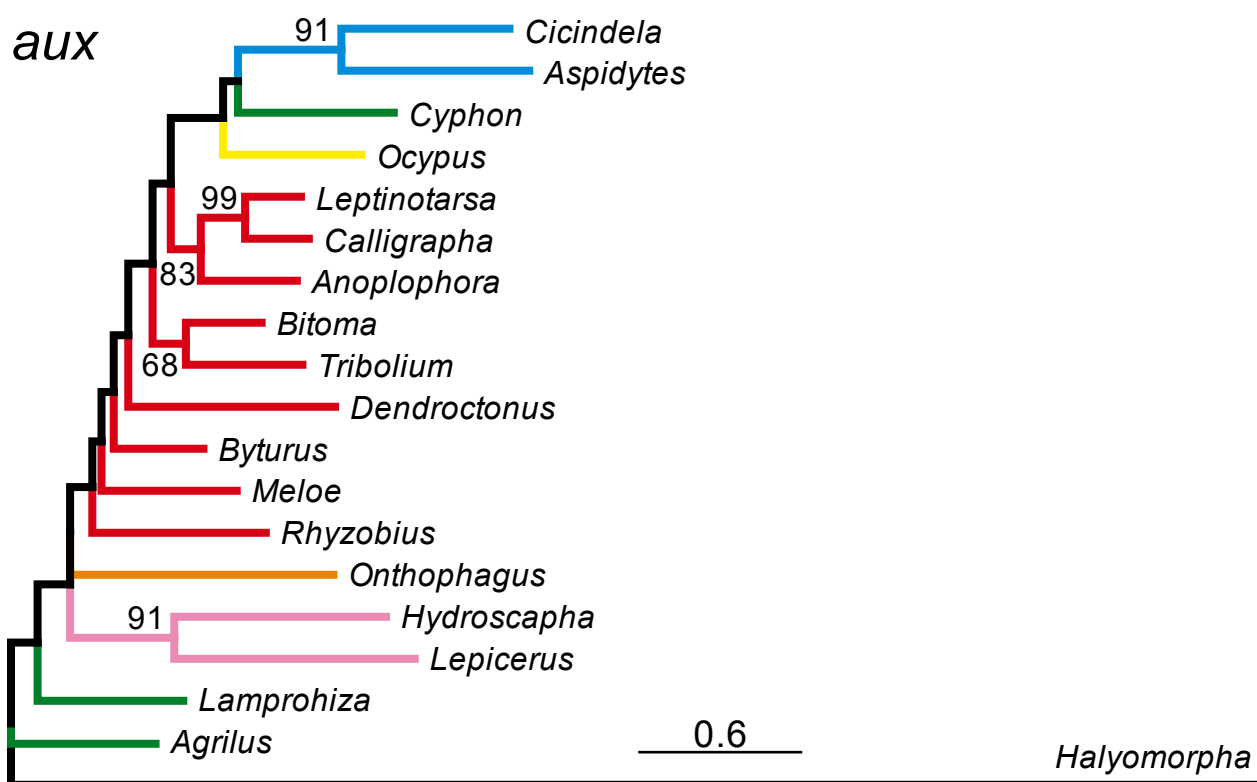
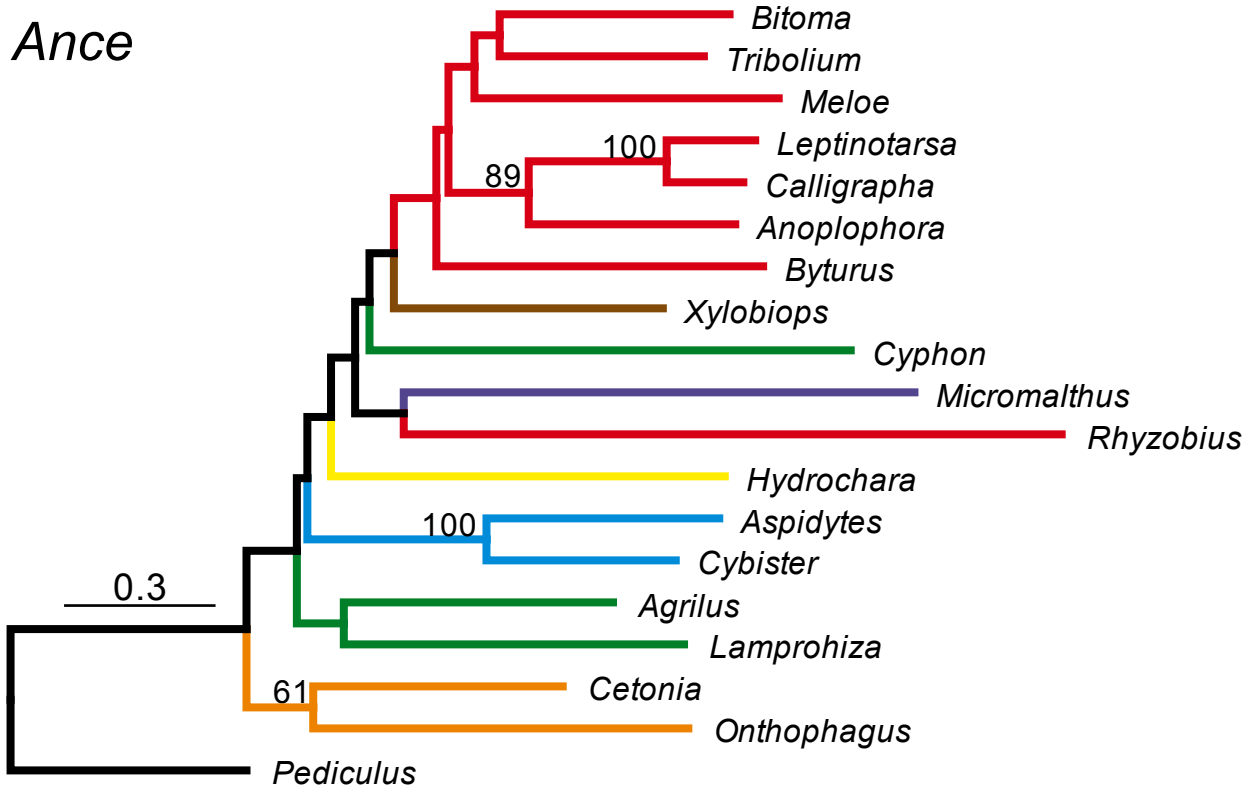
# Vps28



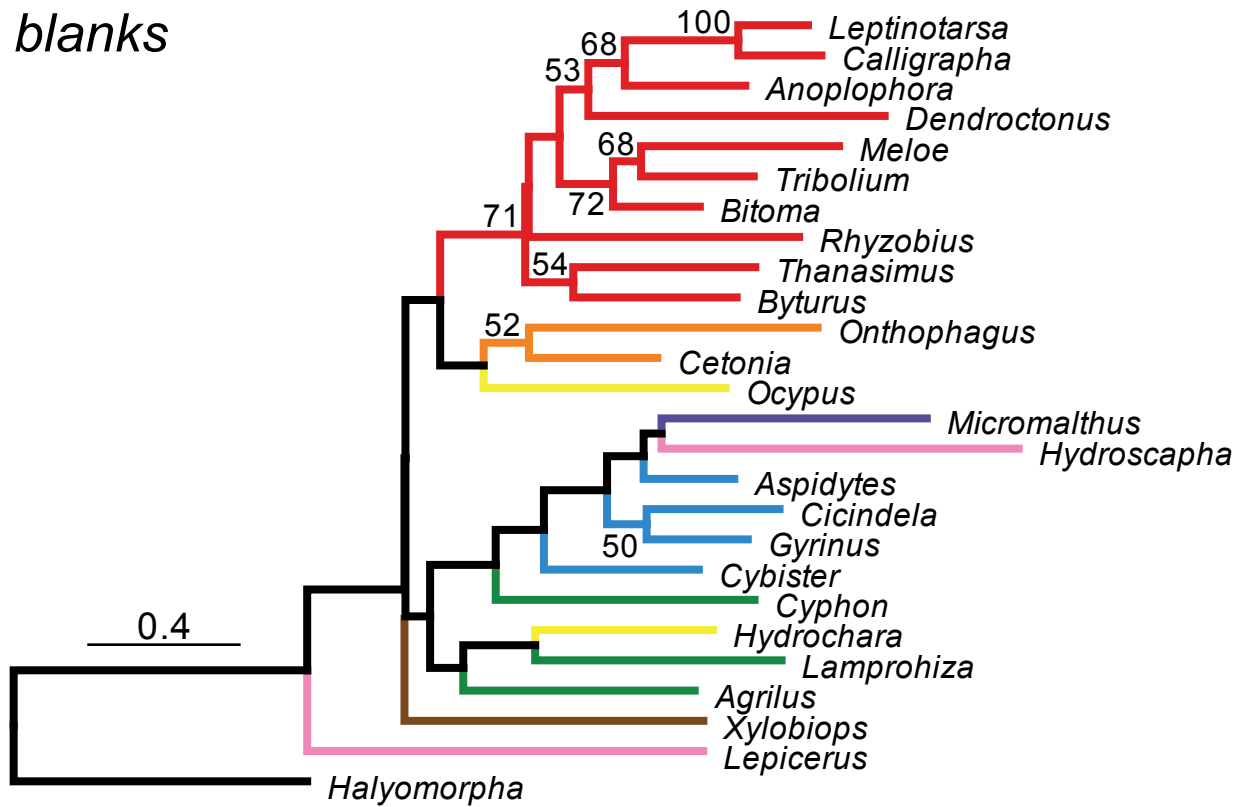
**File S3.** Maximum likelihood trees based on the nucleotide alignments of different sperm individualization genes in beetles.

Tree	Gene	Model	Gamma	Invariant	Log likelihood
1	<i>Ance</i>	GTR	0.660	0.146	-24961.88327
2	<i>aux</i>	GTR	0.386	0.070	-27681.59804
3	<i>blanks</i>	GTR	0.931	0.100	-12607.08759
4	<i>Bug22</i>	GTR	0.222	NA	-3454.61489
5	<i>CdsA</i>	GTR	0.290	0.154	-17968.57896
6	<i>Chc</i>	GTR	0.368	0.379	-55858.11062
7	<i>Ctp</i>	GTR	0.096	NA	-1384.6439
8	<i>Cul3</i>	TN93	0.175	0.204	-28997.57548
9	<i>Dark</i>	GTR	1.265	0.092	-20388.14695
10	<i>didum</i>	GTR	0.759	0.209	-60072.64694
11	<i>Dredd</i>	GTR	1.197	0.096	-14618.72638
12	<i>Dronc</i>	GTR	1.355	0.099	-15698.08633
13	<i>Duba</i>	GTR	0.527	0.176	-12010.48828
14	<i>EcR</i>	GTR	0.449	0.365	-12199.45136
15	<i>eIF3m</i>	GTR	0.329	0.137	-18855.00527
16	<i>Fadd</i>	GTR	1.321	0.084	-4356.88960
17	<i>gish</i>	GTR	0.211	0.170	-15923.19717
18	<i>gudu</i>	GTR	0.618	0.135	-22338.05316
19	<i>heph</i>	TN93	0.396	0.275	-8479.41719
20	<i>hmw</i>	GTR	0.898	NA	-1707.47296
21	<i>jar</i>	GTR	0.708	0.241	-52172.98272
22	<i>klhl10</i>	GTR	0.504	0.239	-17745.77407
23	<i>Lasp</i>	GTR	0.474	0.351	-5407.13219
24	<i>Mer</i>	GTR	0.598	0.356	-22473.10341
25	<i>mlt</i>	GTR	0.593	0.196	-19352.46023
26	<i>nes</i>	GTR	0.911	0.148	-25494.50312
27	<i>Npc1a</i>	GTR	0.724	0.204	-58520.58794
28	<i>nsr</i>	GTR	0.287	0.143	-11564.48392
29	<i>orb2</i>	GTR	0.327	0.450	-4091.42072
30	<i>Osbp</i>	GTR	0.699	0.244	-29993.57419
31	<i>oys</i>	GTR	0.773	0.158	-15795.25364
32	<i>Past1</i>	GTR	0.474	0.325	-14648.80870
33	<i>Pen</i>	GTR	0.676	0.232	-26388.05898
34	<i>poe</i>	GTR	0.590	0.253	-50399.32806
35	<i>porin</i>	GTR	0.674	0.191	-15348.95453
36	<i>Prosalpha6T</i>	GTR	0.659	0.312	-13745.91664
37	<i>scat</i>	GTR	0.556	0.147	-34071.26981
38	<i>shi</i>	GTR	0.552	0.395	-34985.87502
39	<i>skap</i>	GTR	0.596	0.317	-20864.74113
40	<i>sw</i>	GTR	0.473	0.265	-21539.32783
41	<i>Taz</i>	GTR	0.717	0.284	-11982.00743
42	<i>Vps28</i>	GTR	0.241	0.154	-7981.57649

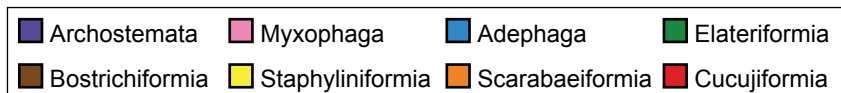
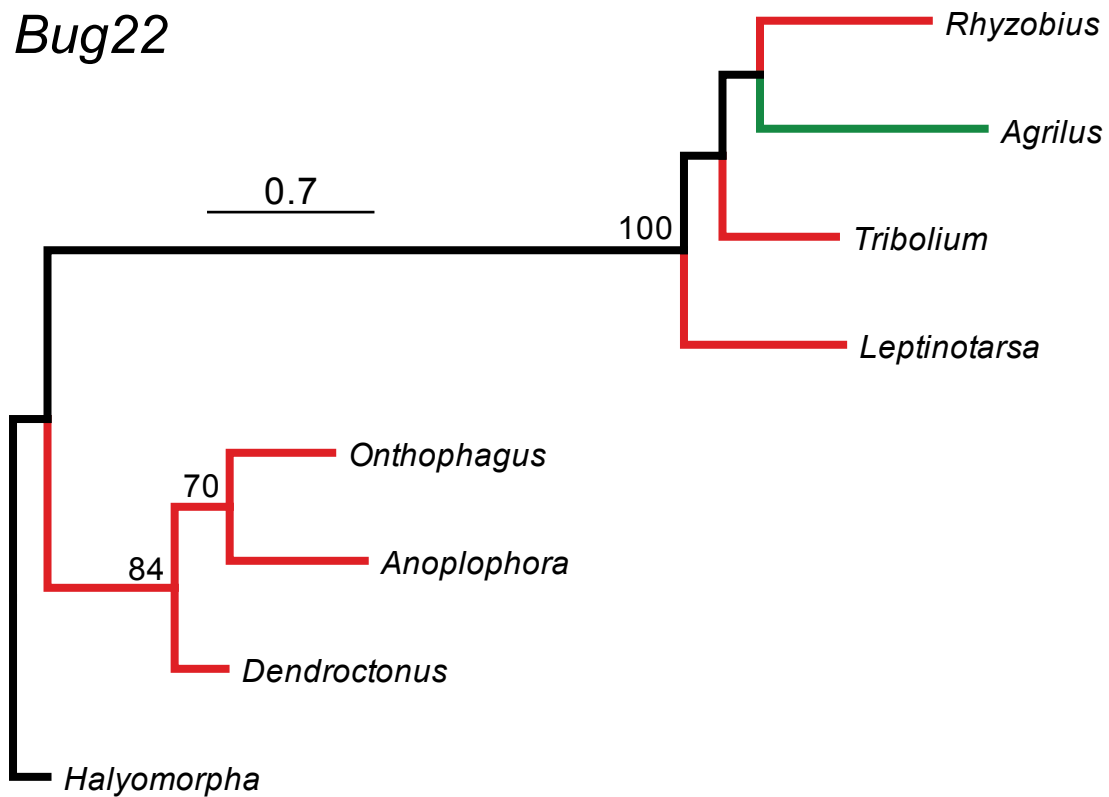




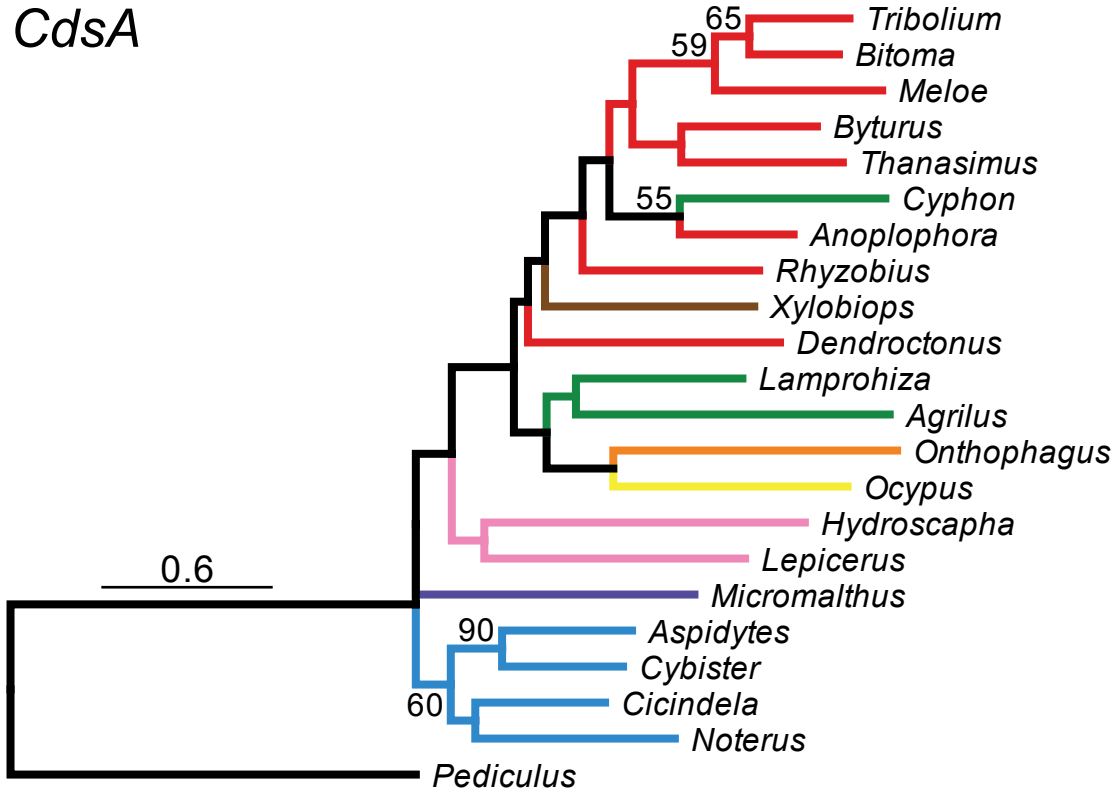
blanks



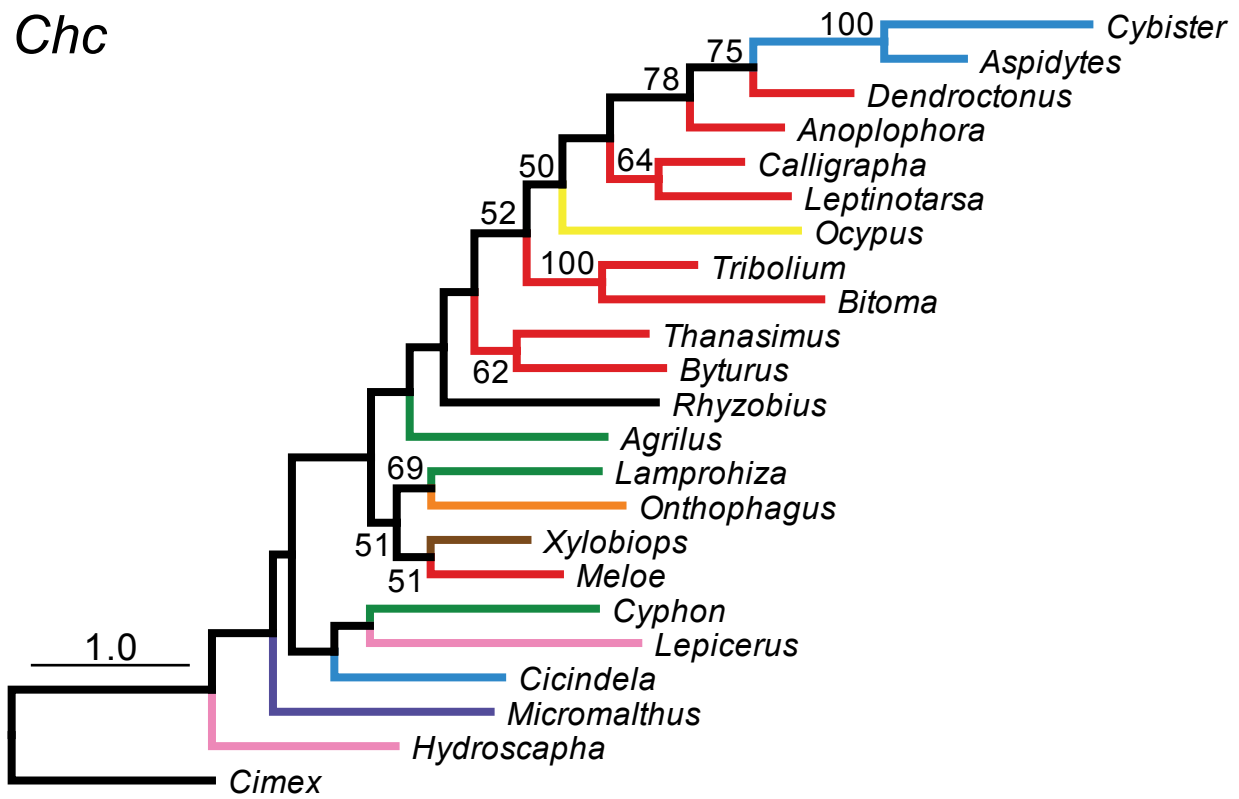
Bug22



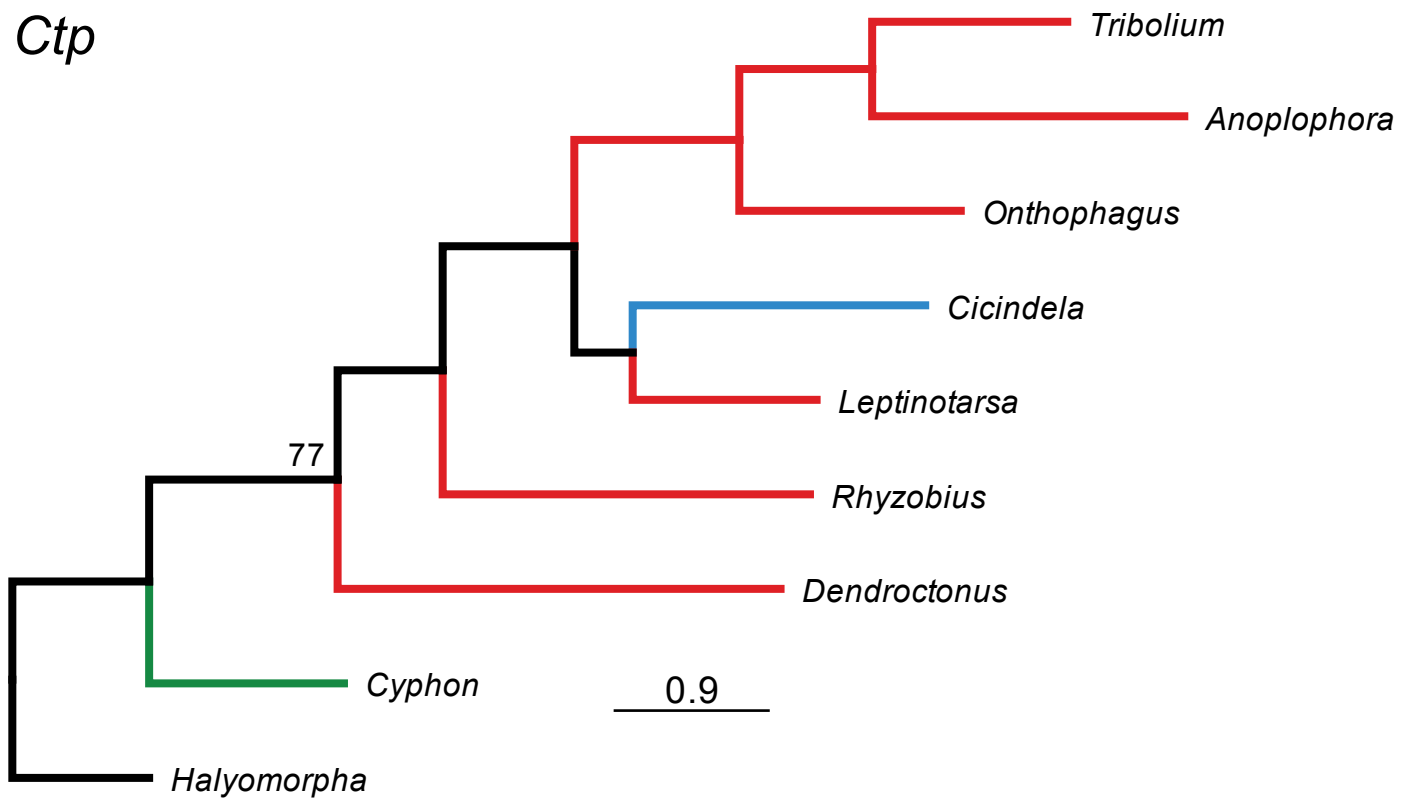
# CdsA



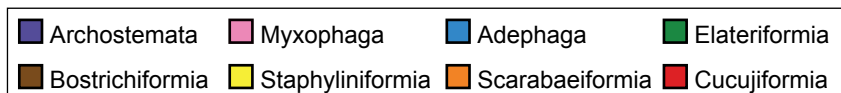
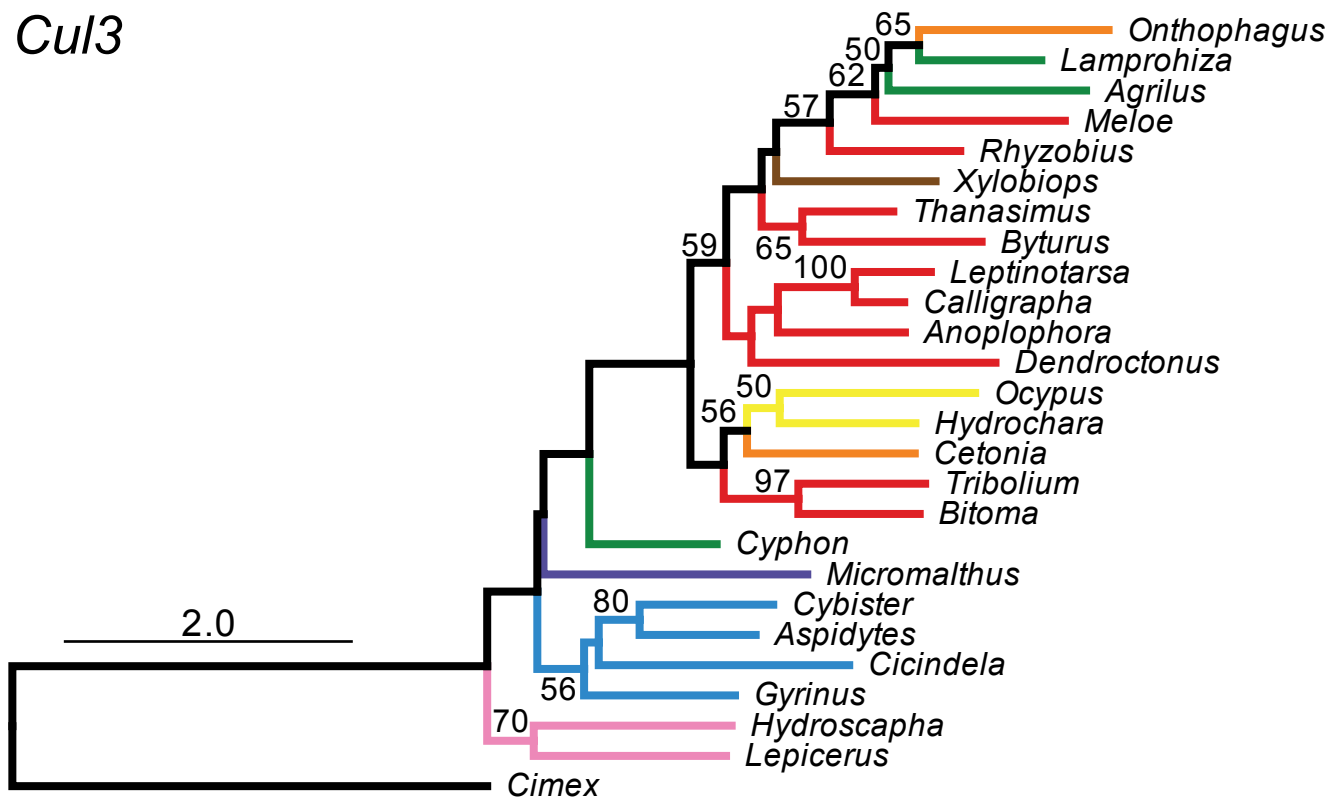
# Chc



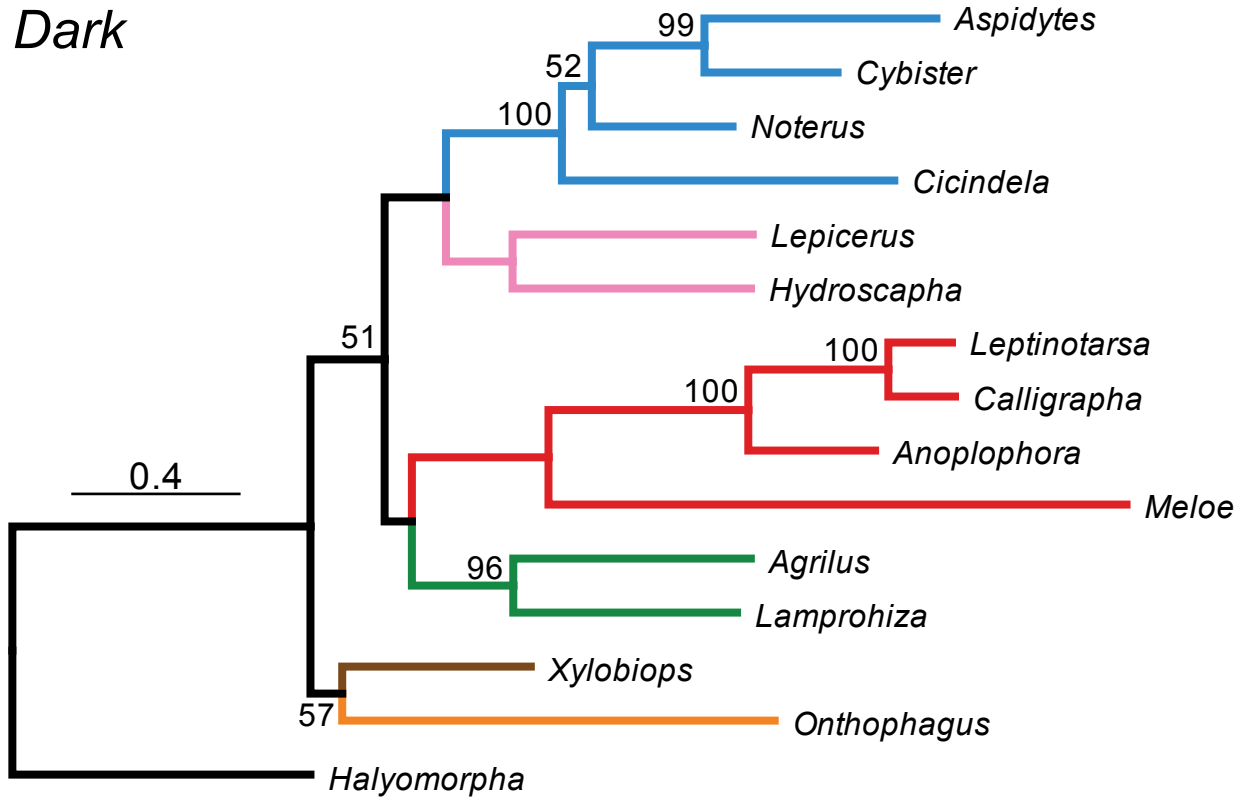
Ctp



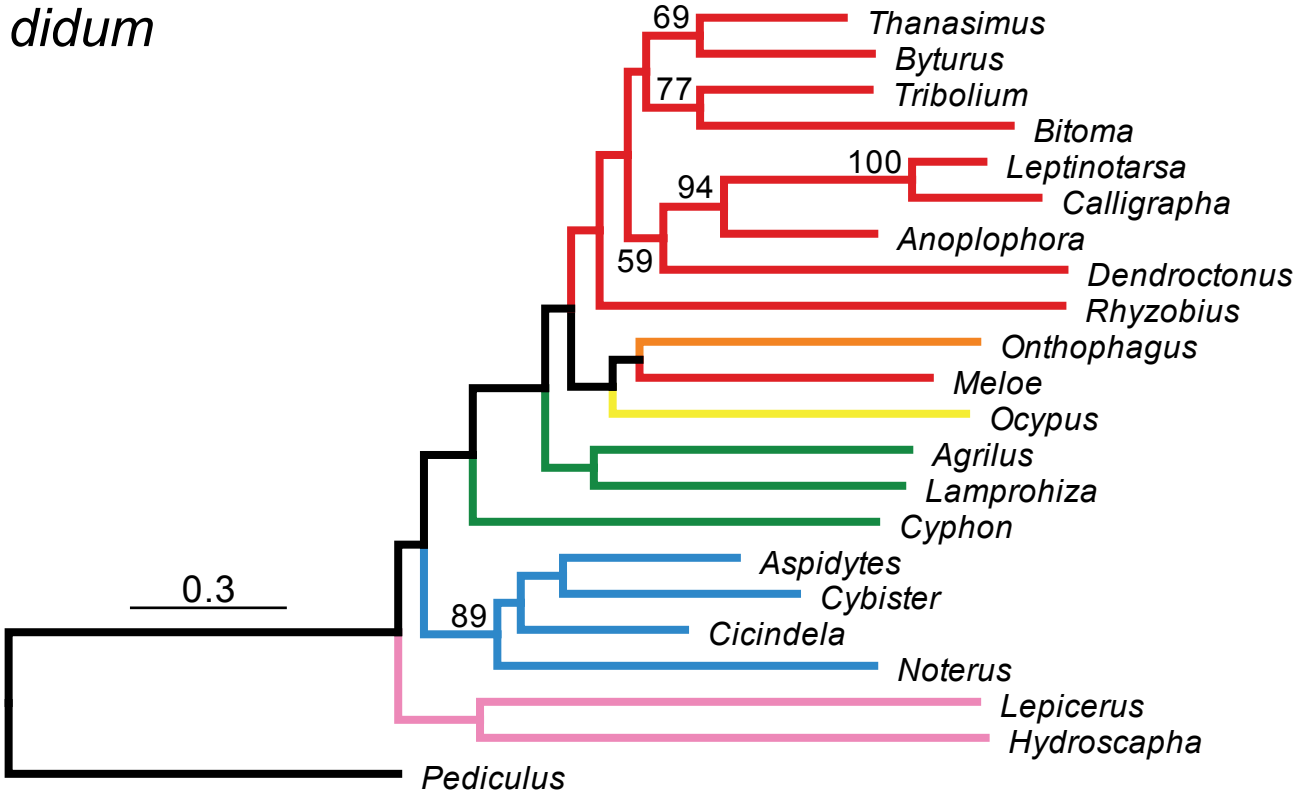
Cul3



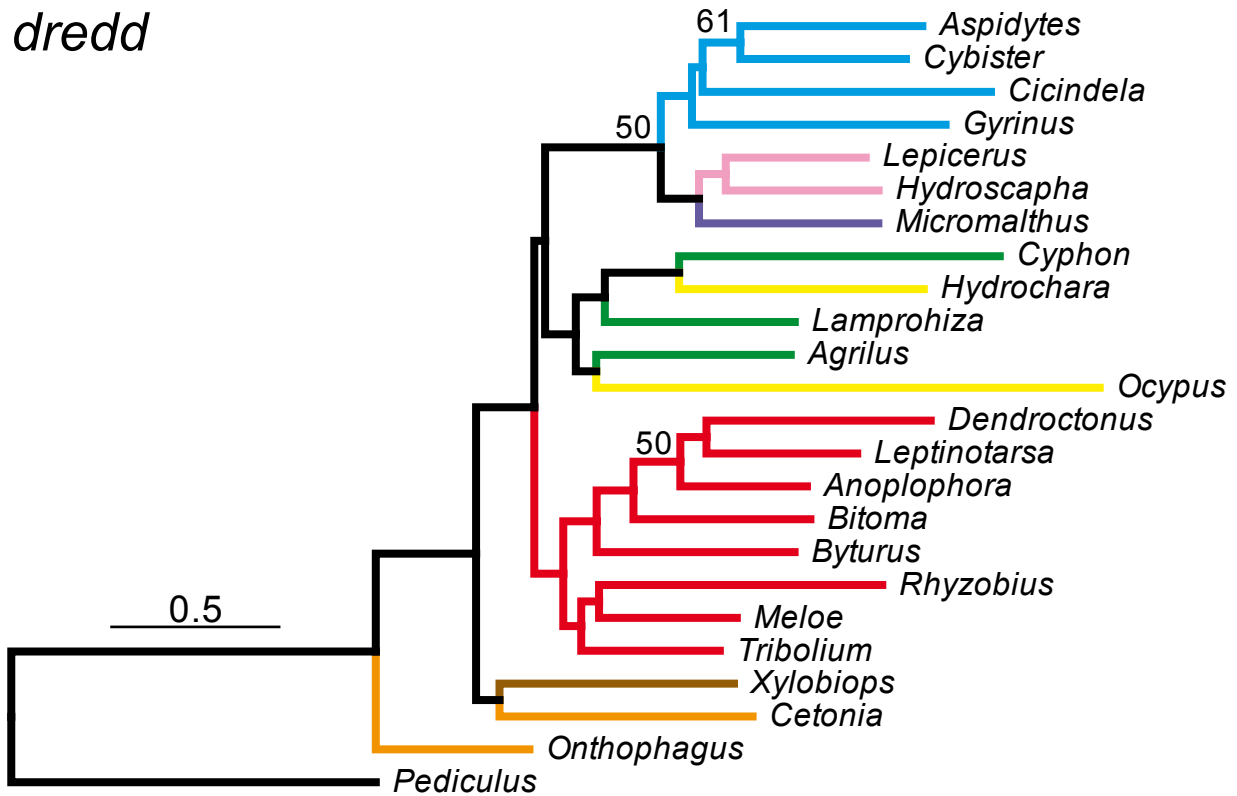
Dark



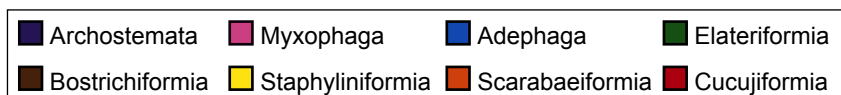
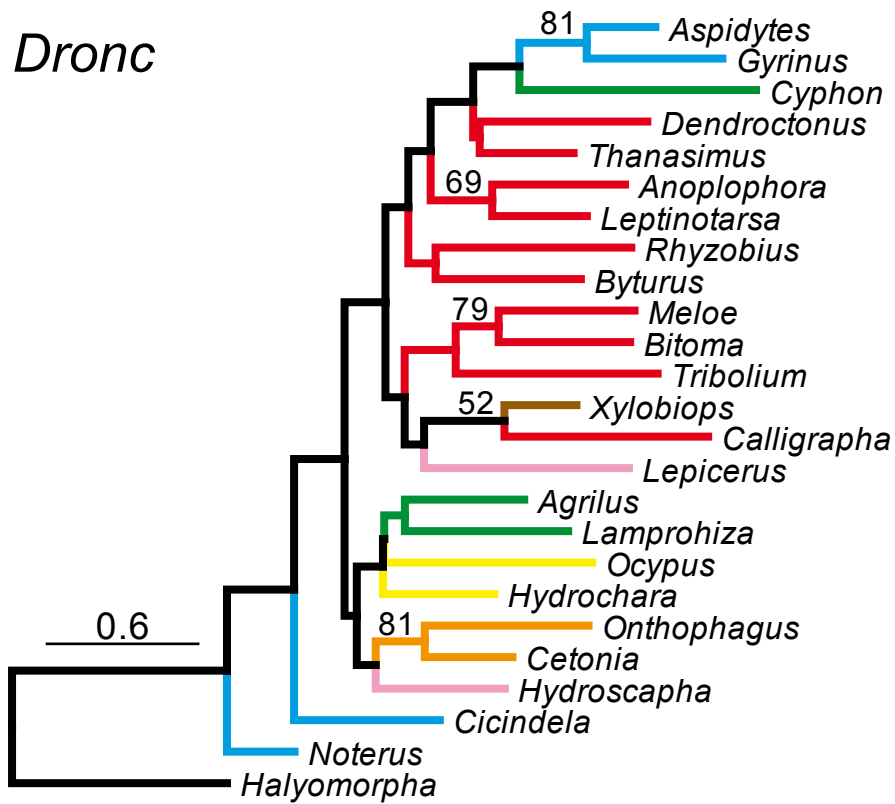
didum



*dredd*

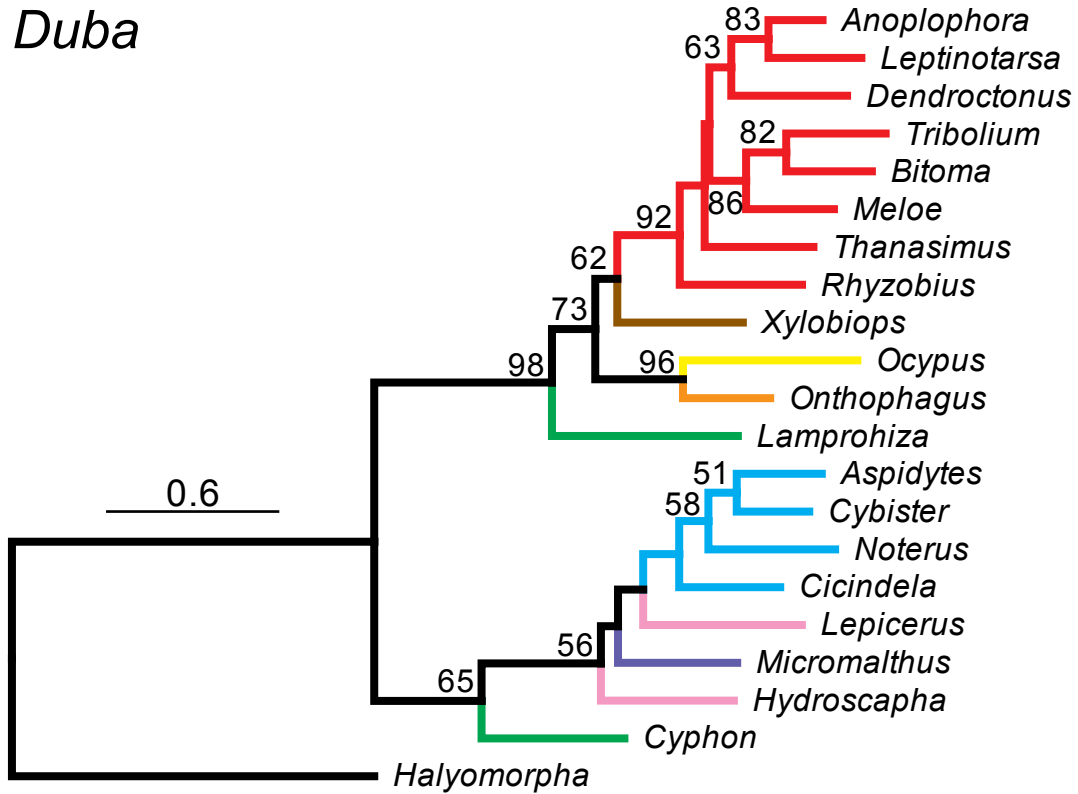


*Dronc*

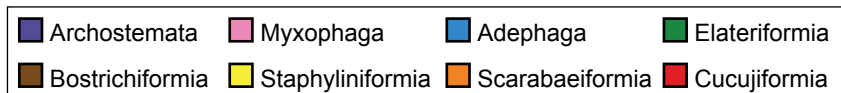
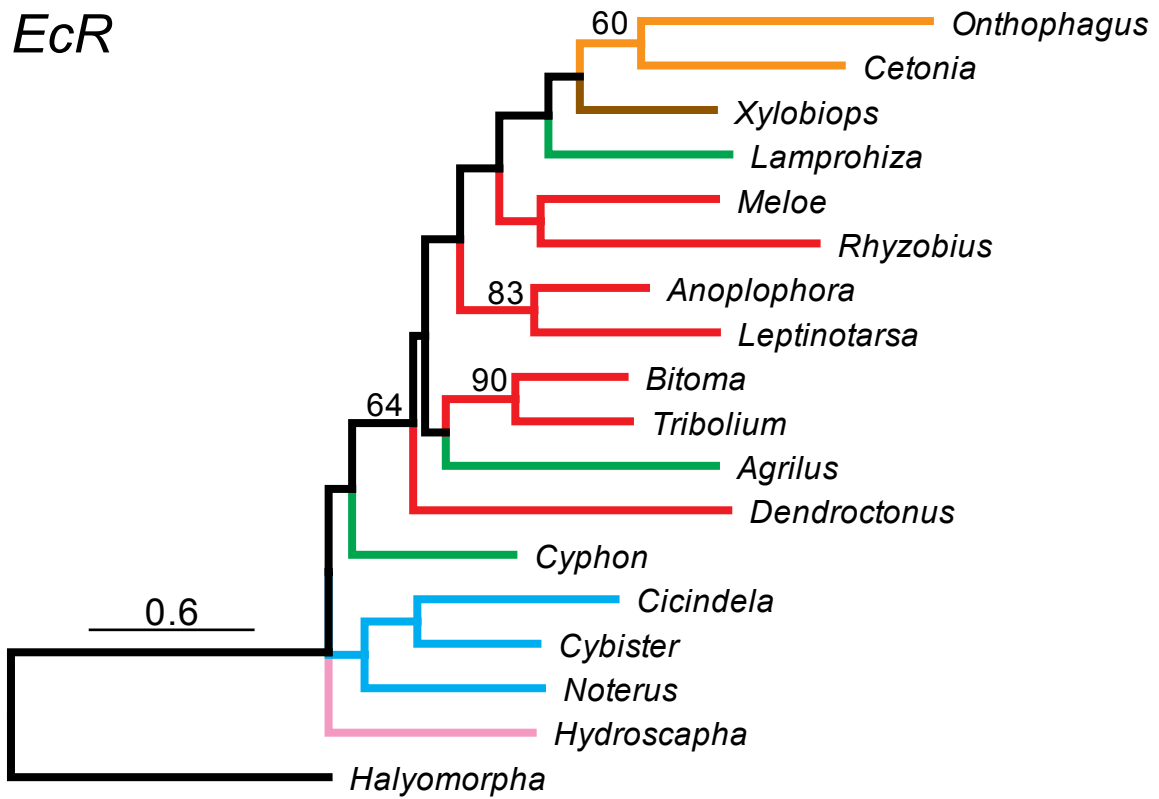




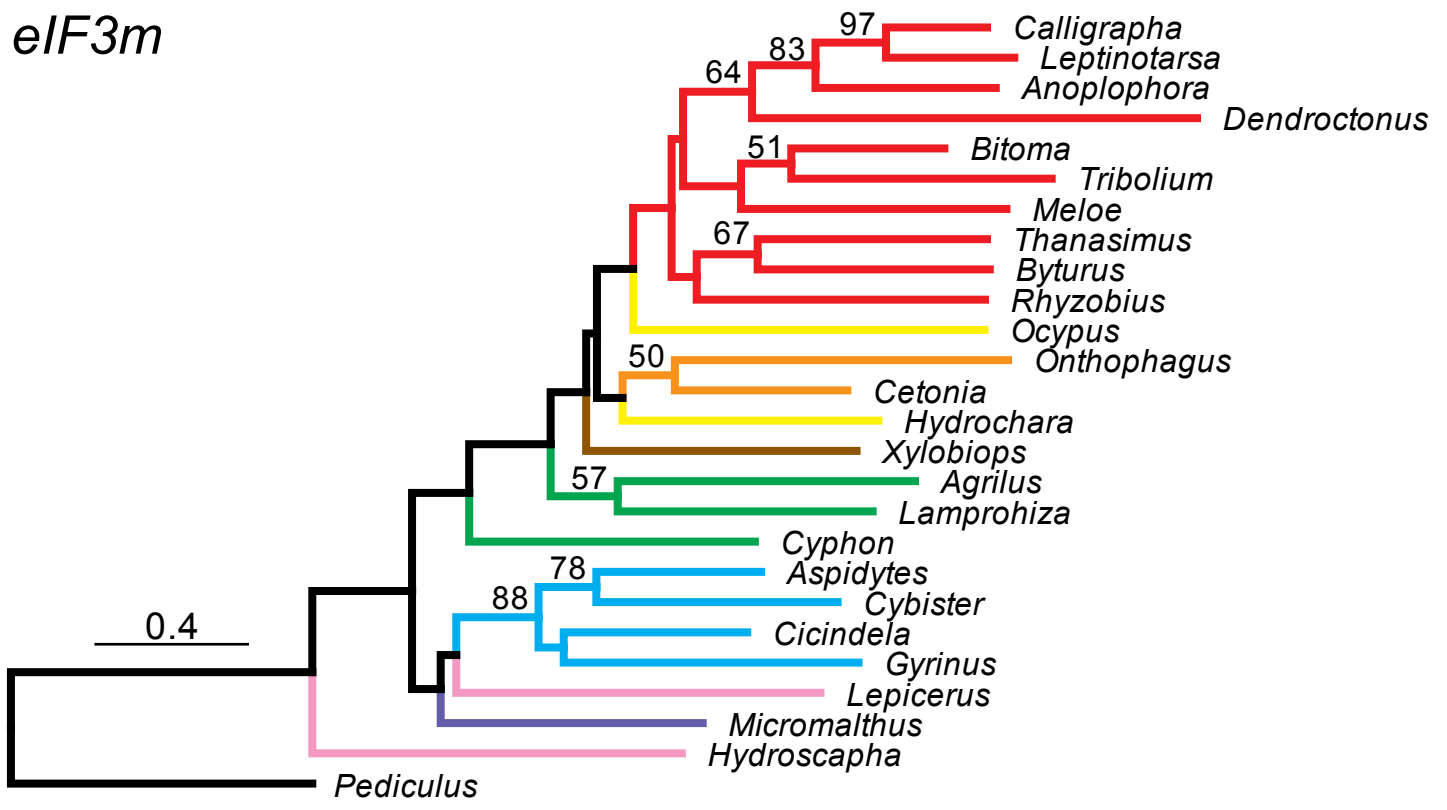
# Duba



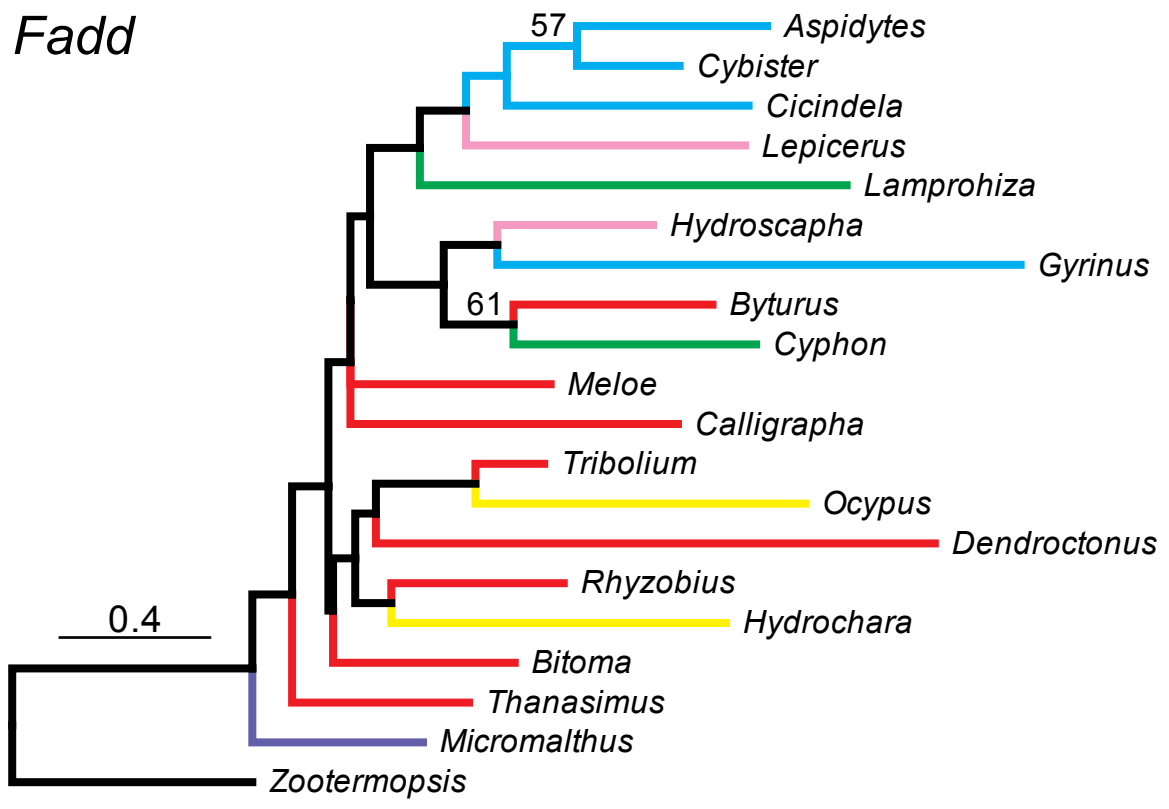
# EcR

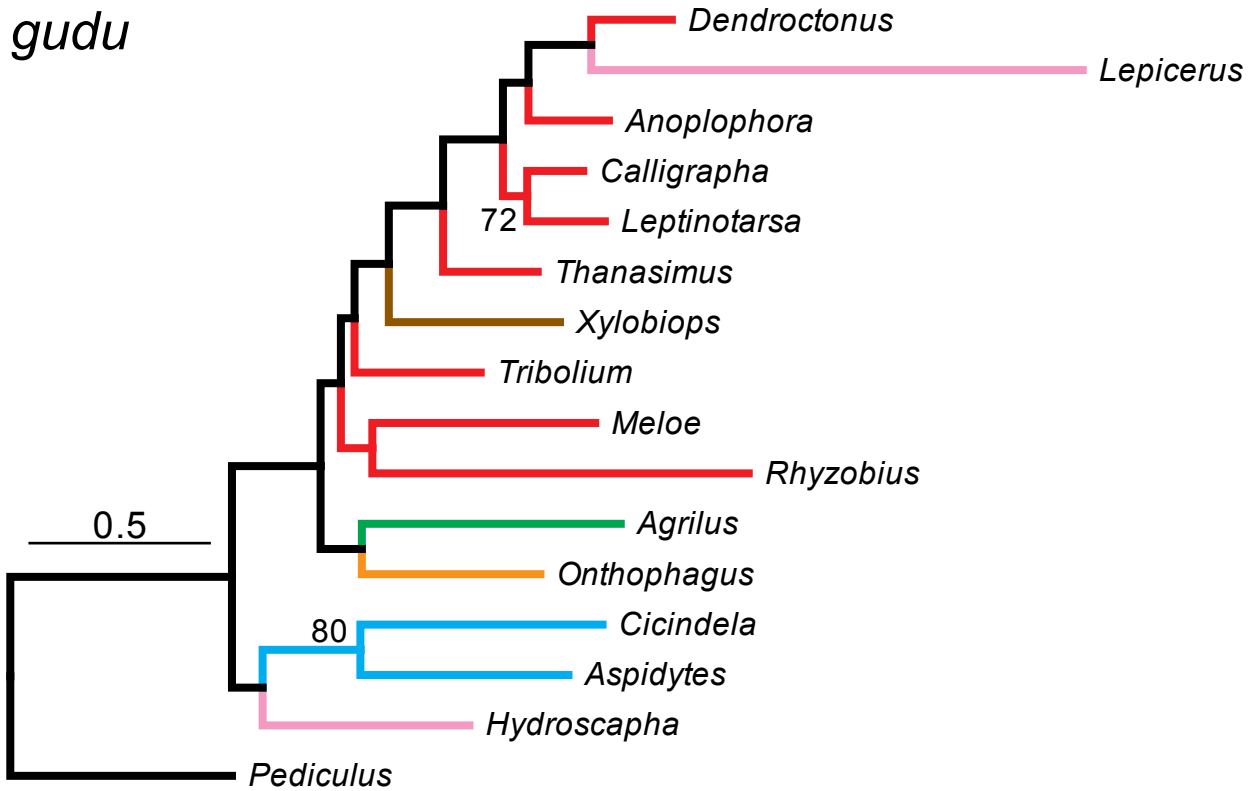
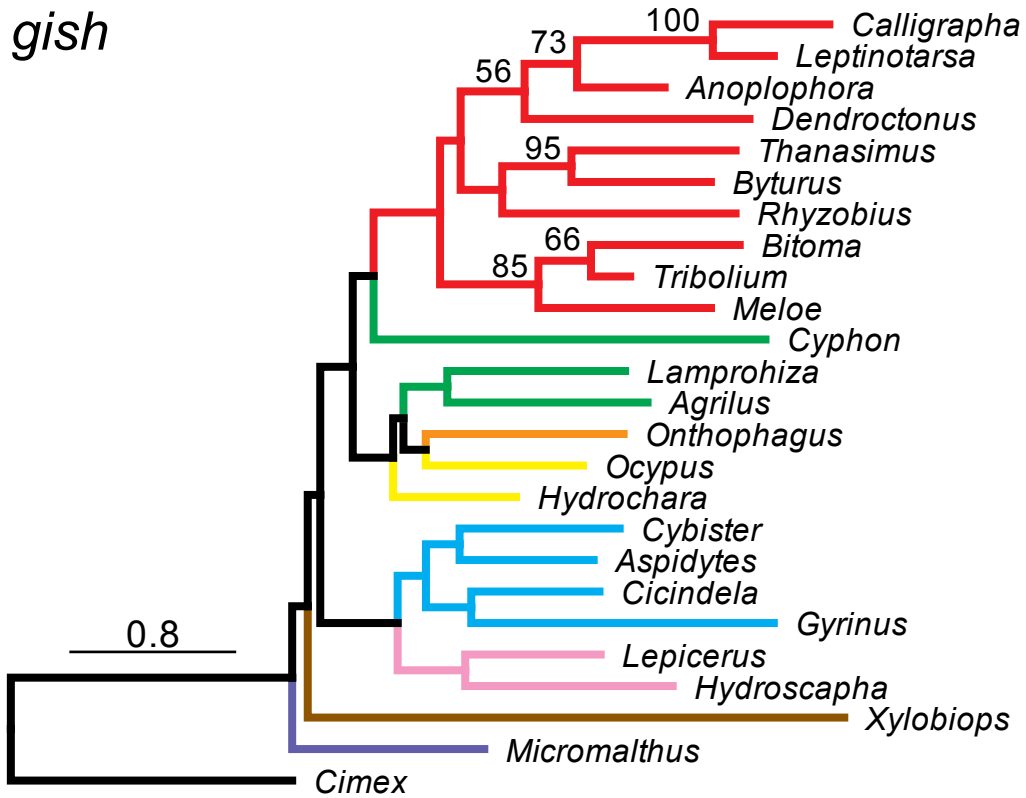


*eIF3m*

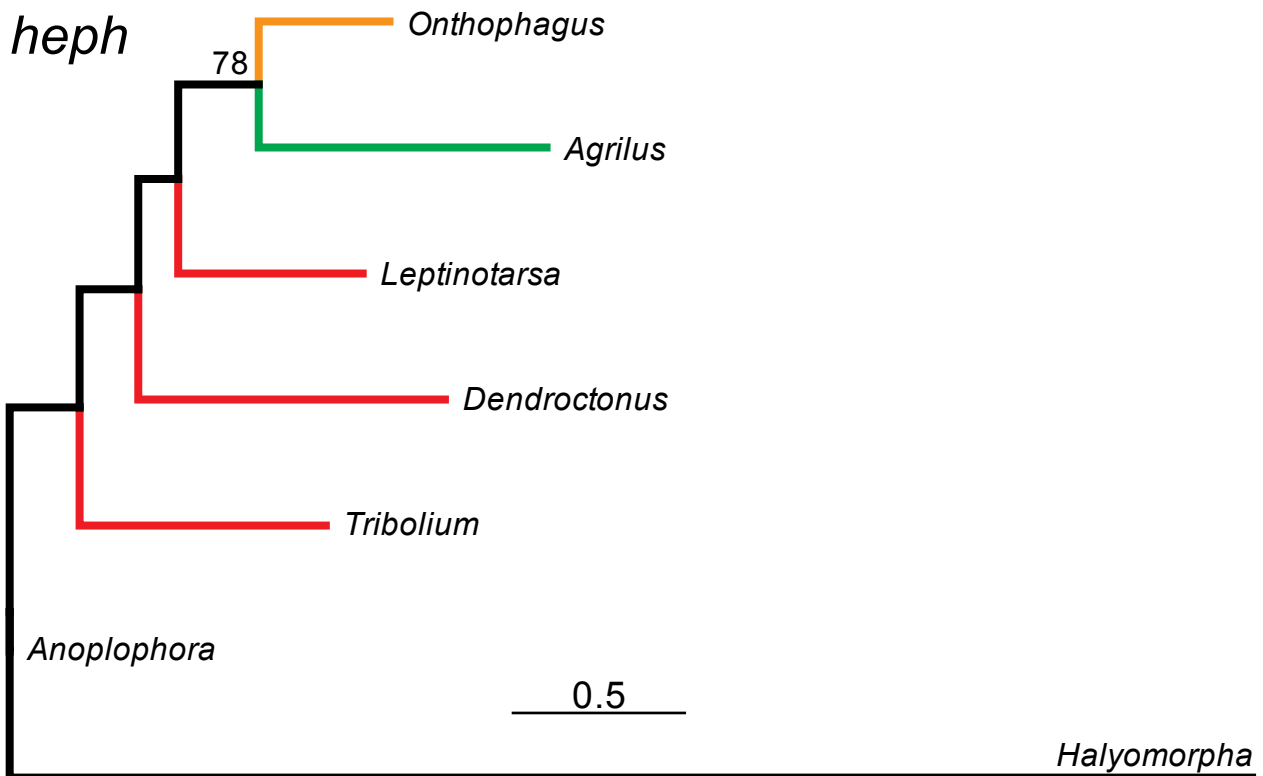


*Fadd*

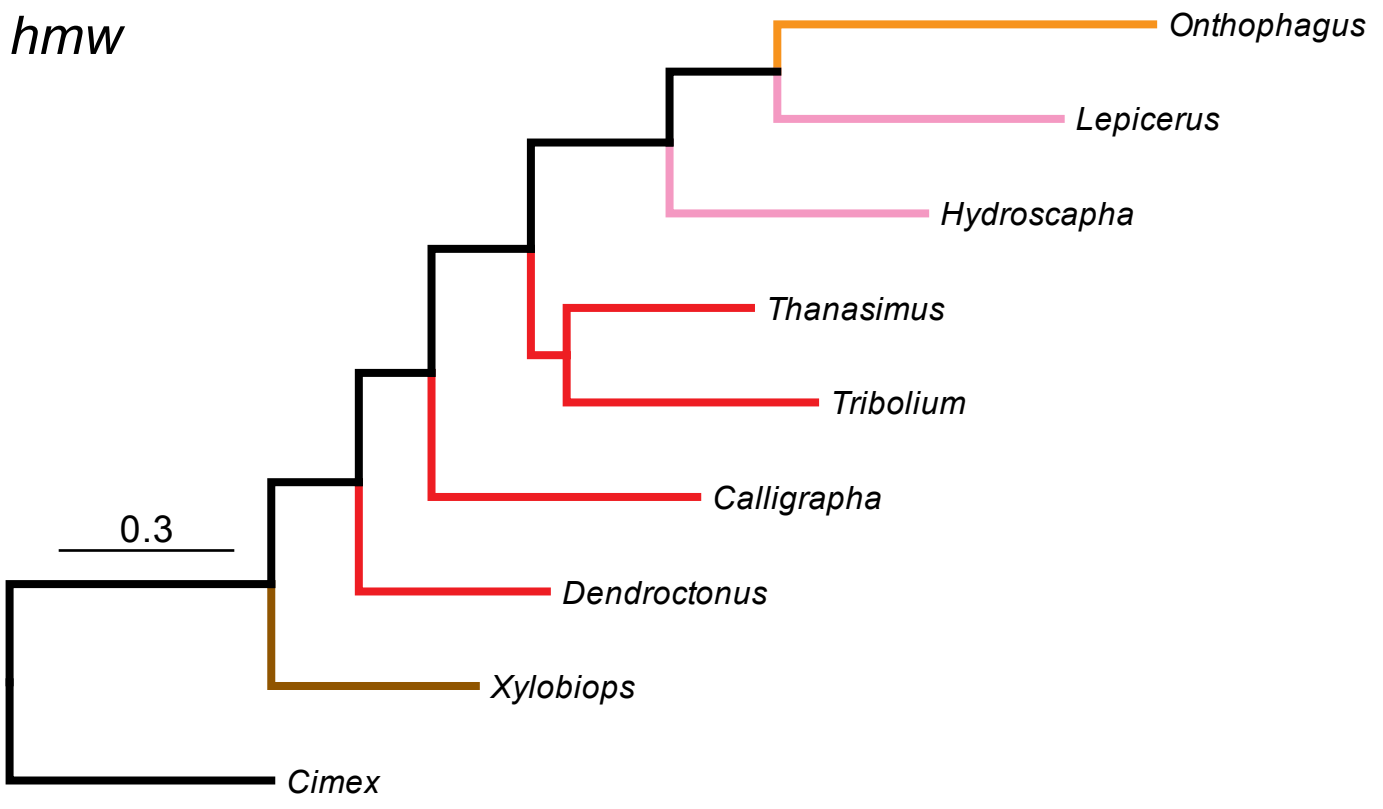




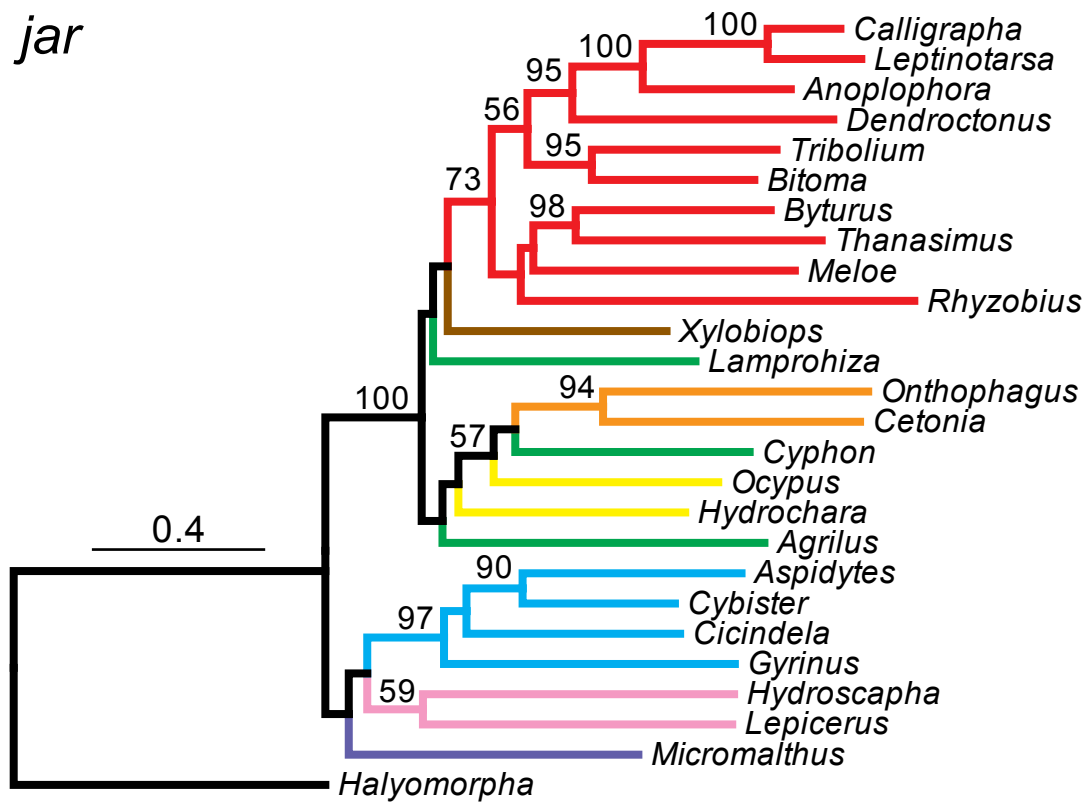
heph



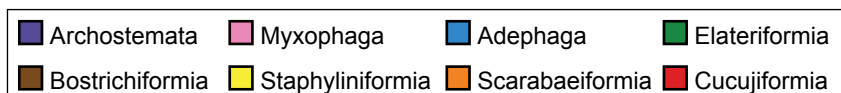
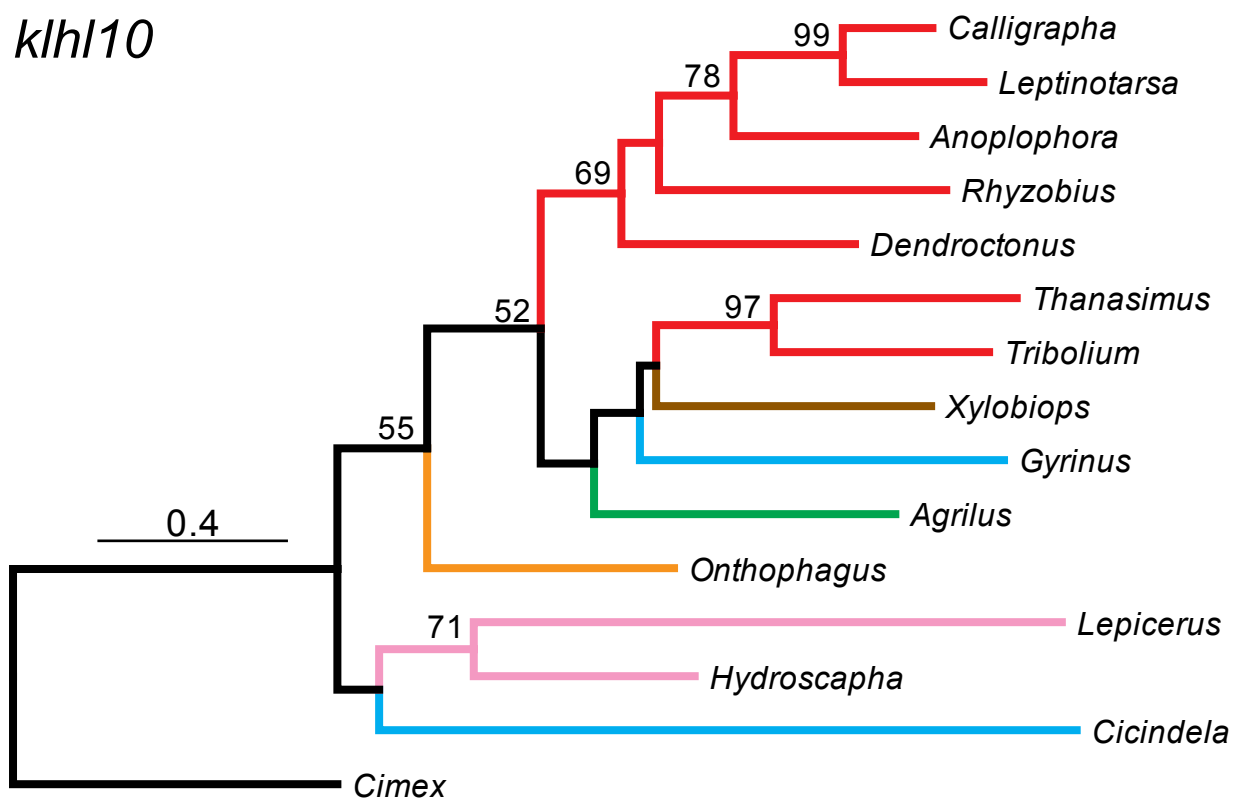
hmv



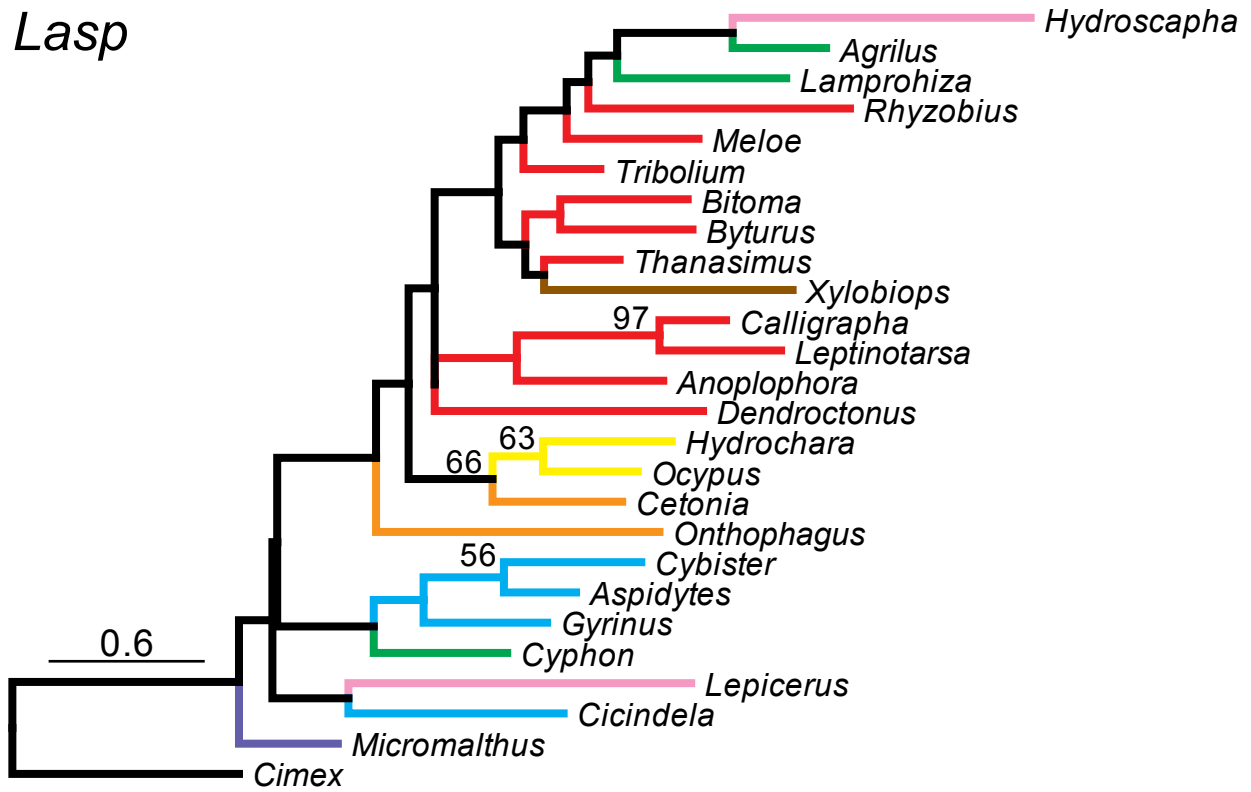
*jar*



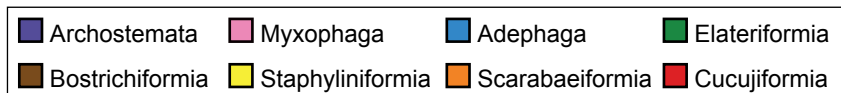
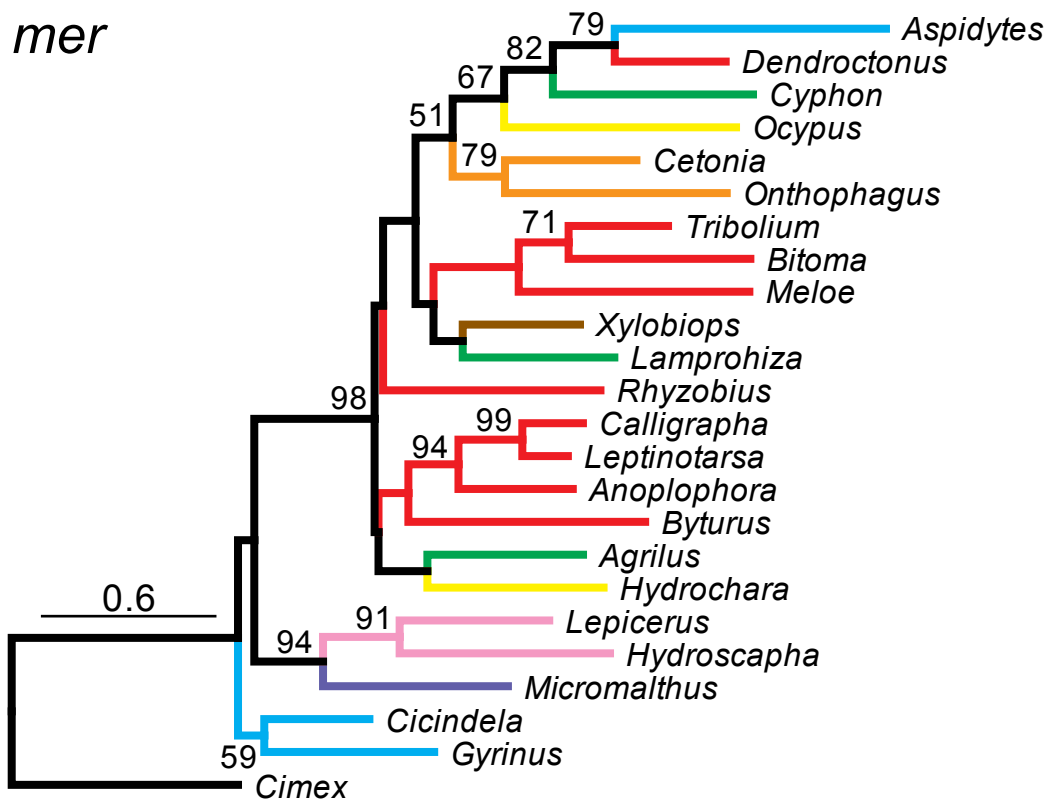
*klhl10*



Lasp

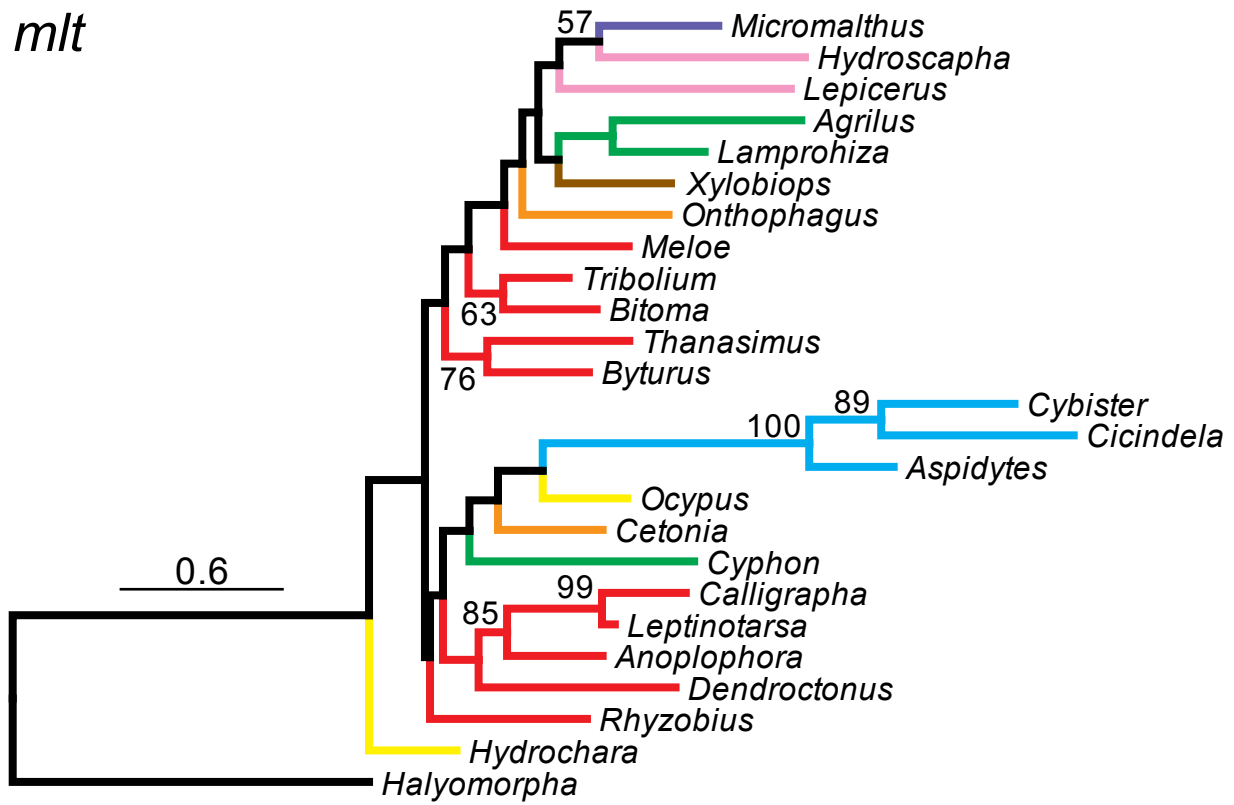


mer

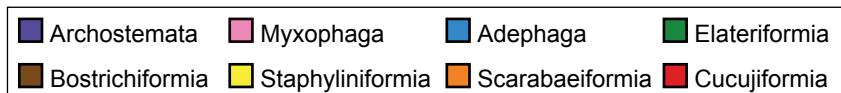
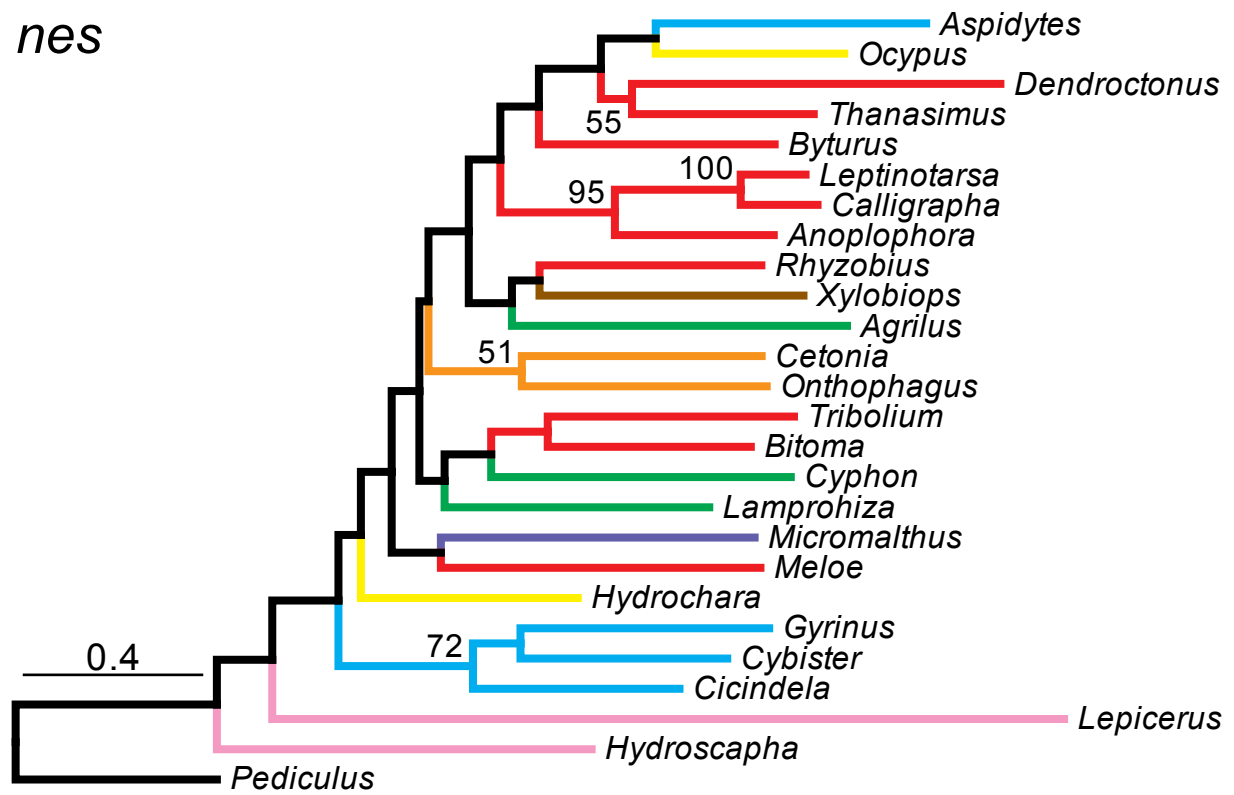




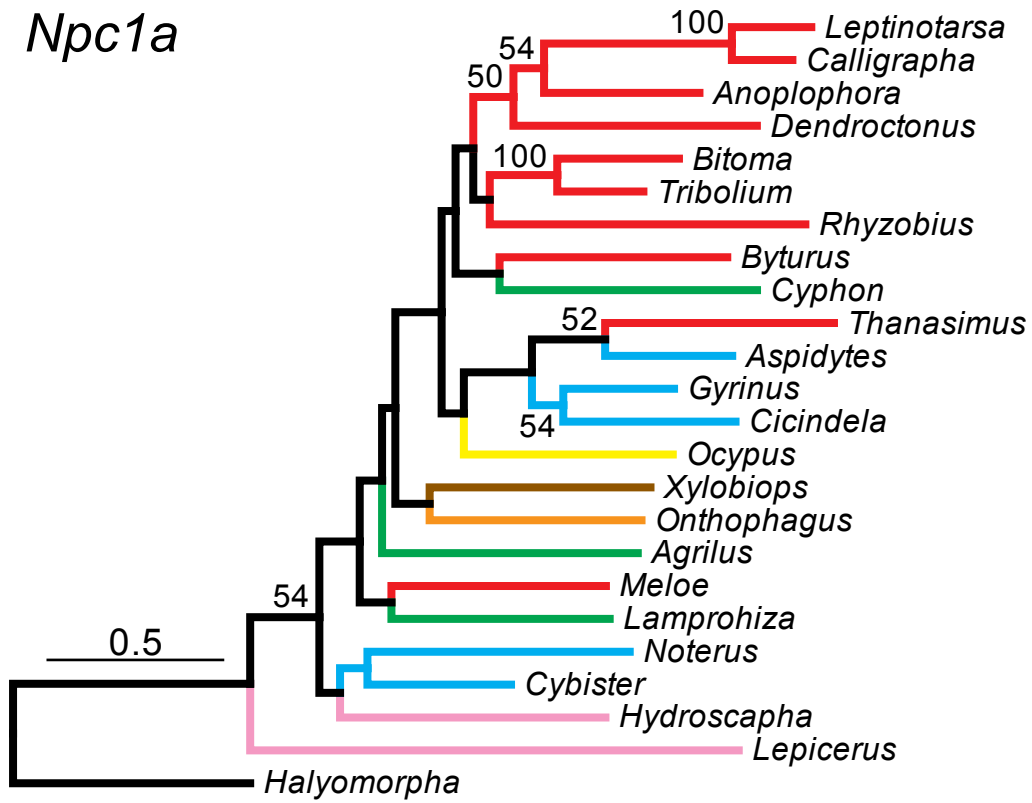
mlt



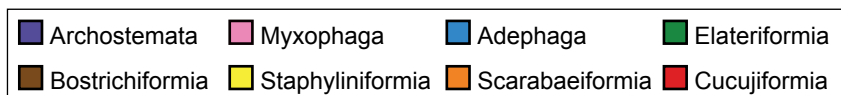
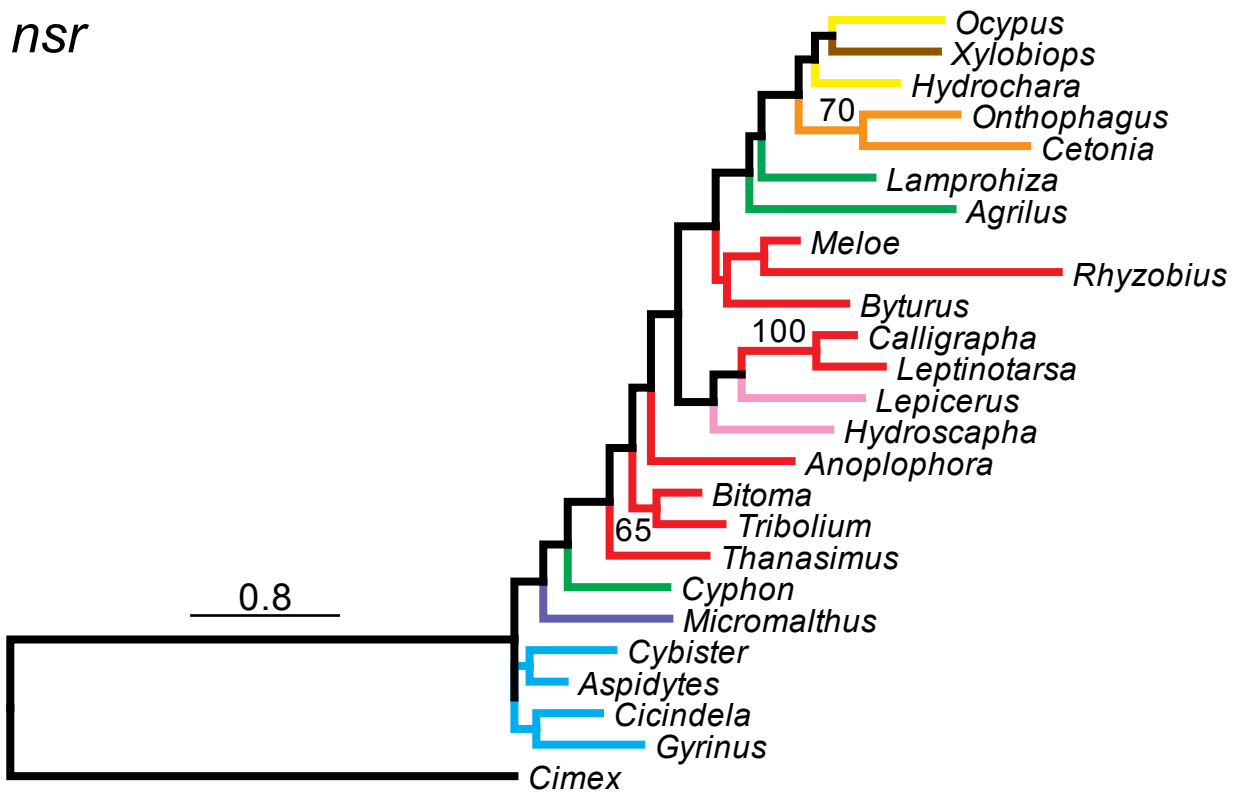
nes



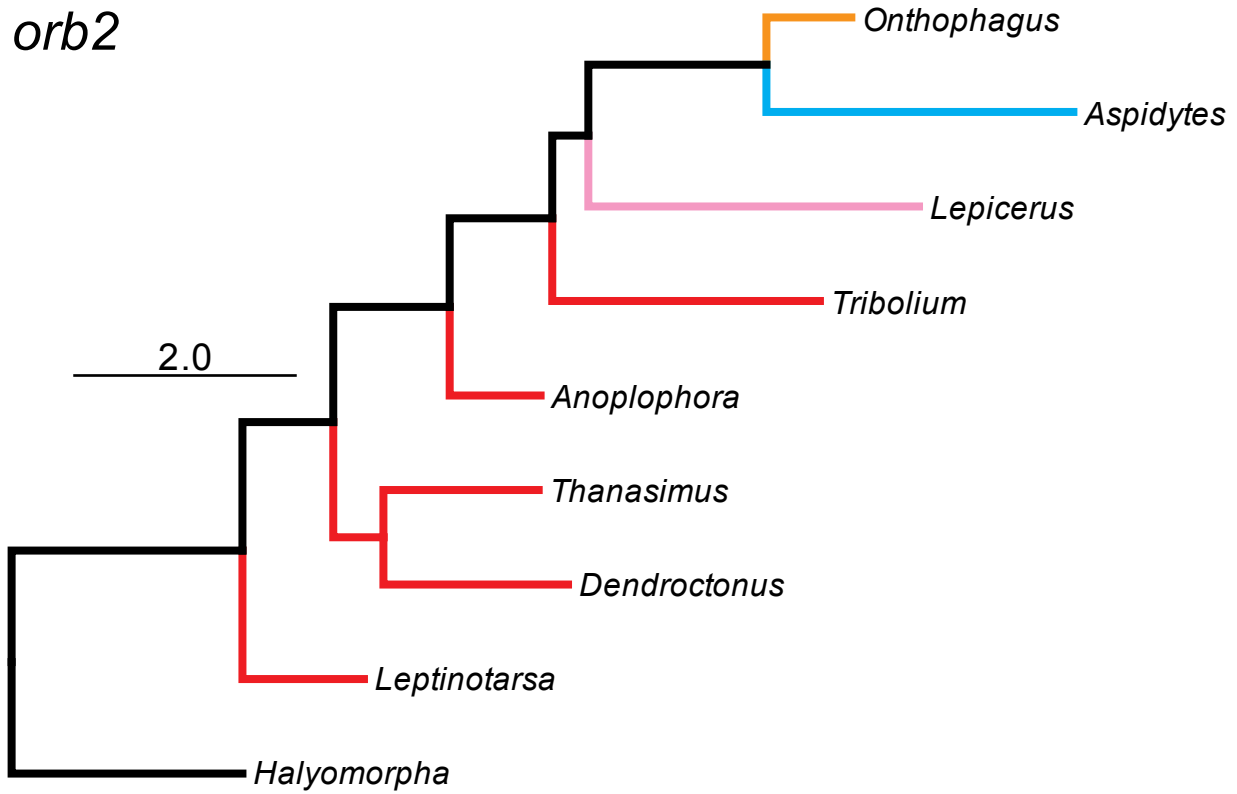
*Npc1a*



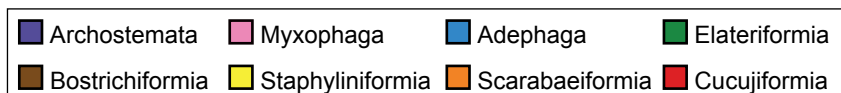
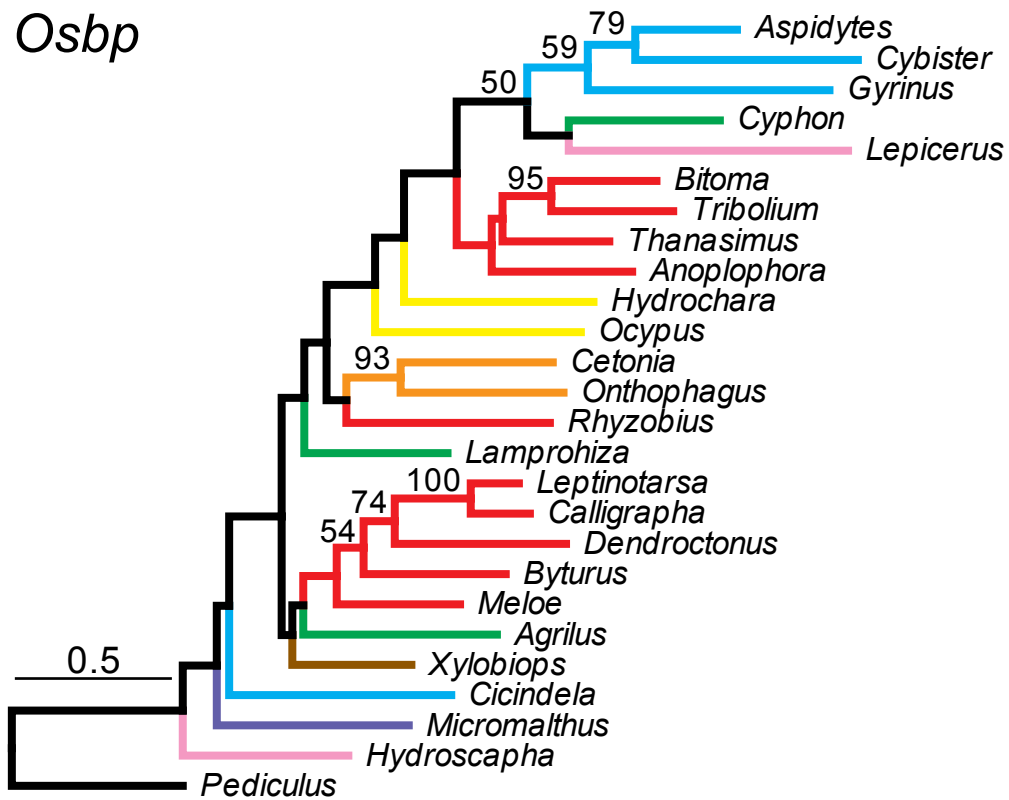
*nsr*



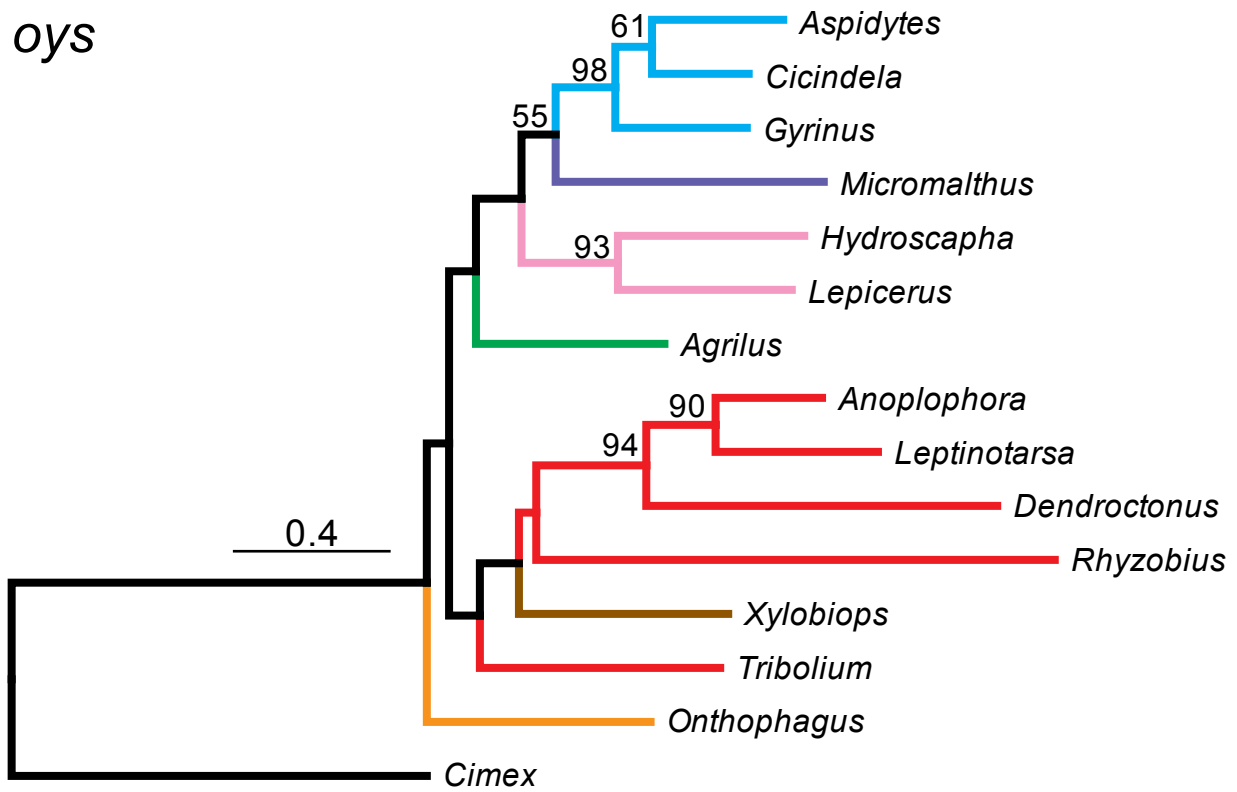
orb2



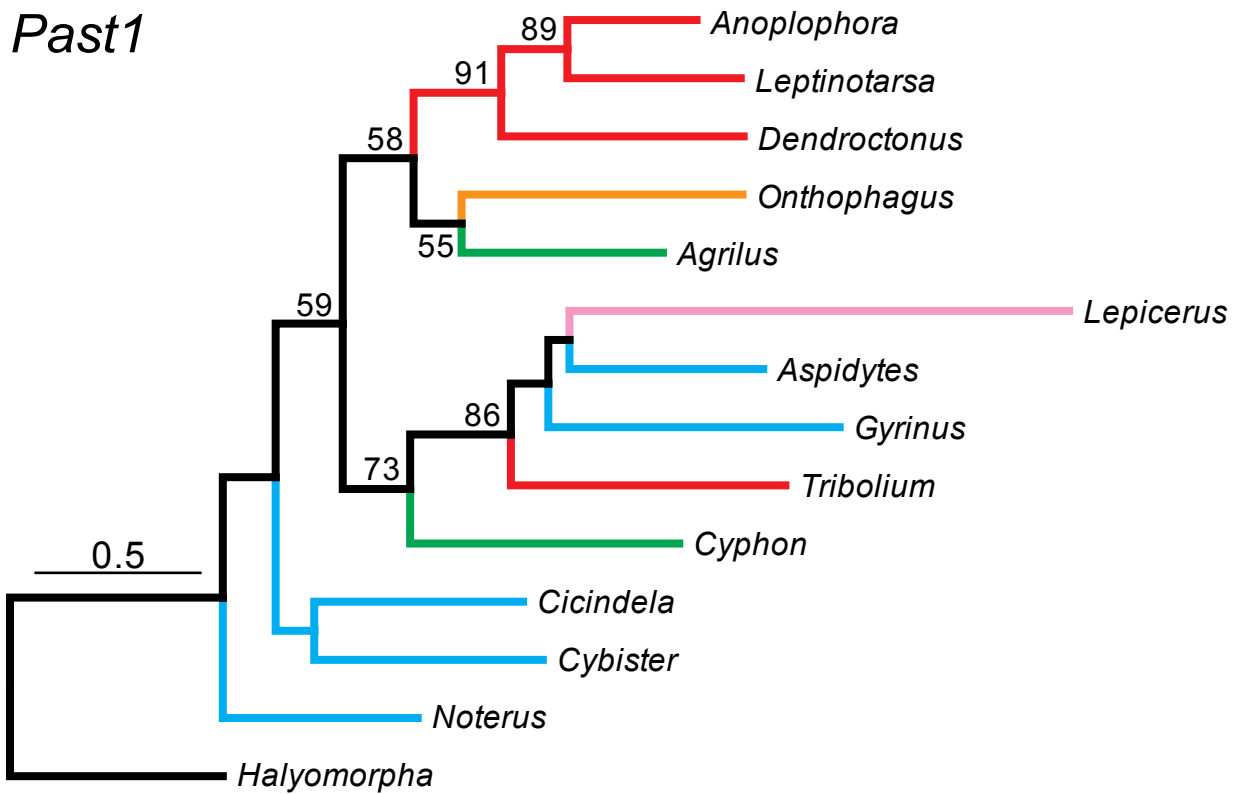
Osbp



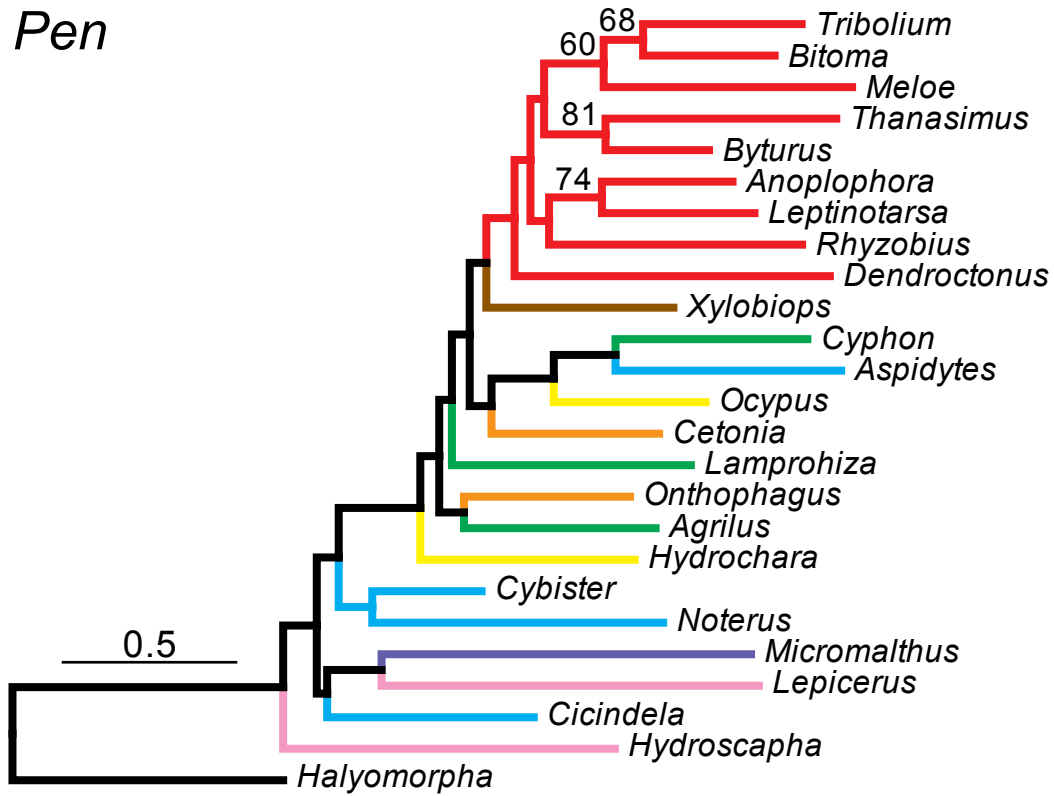
*Oys*



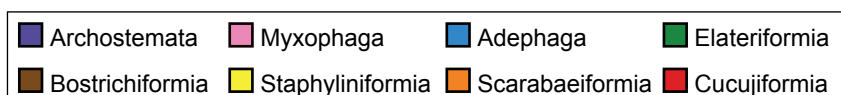
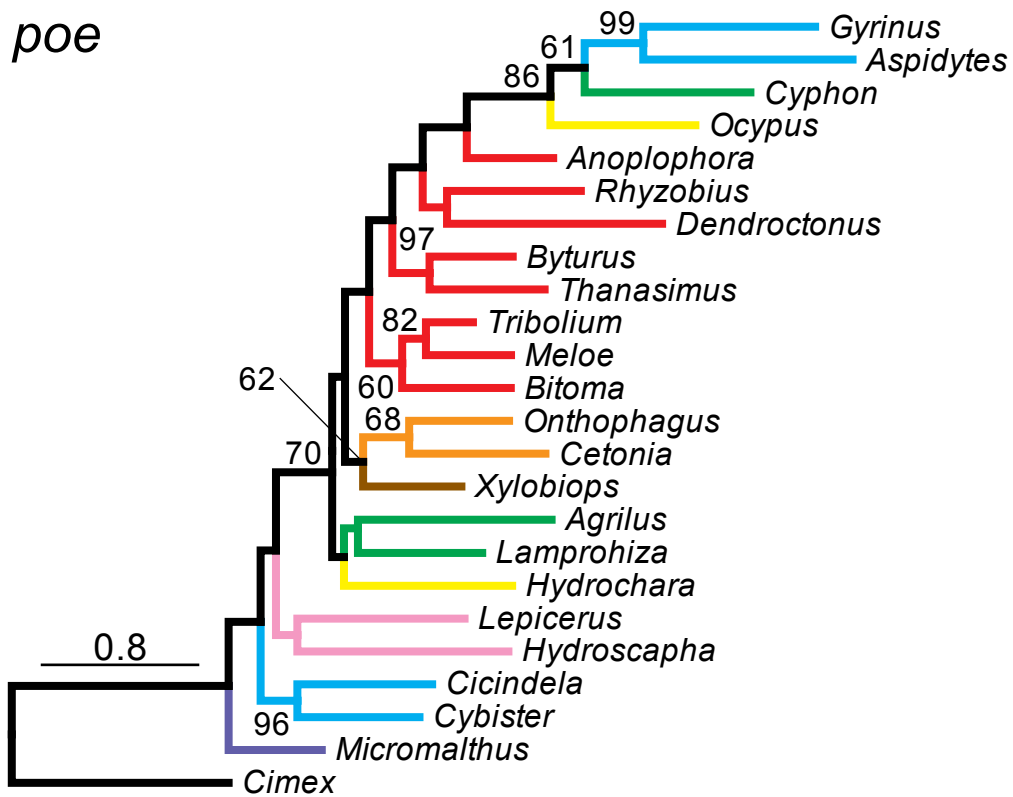
*Past1*



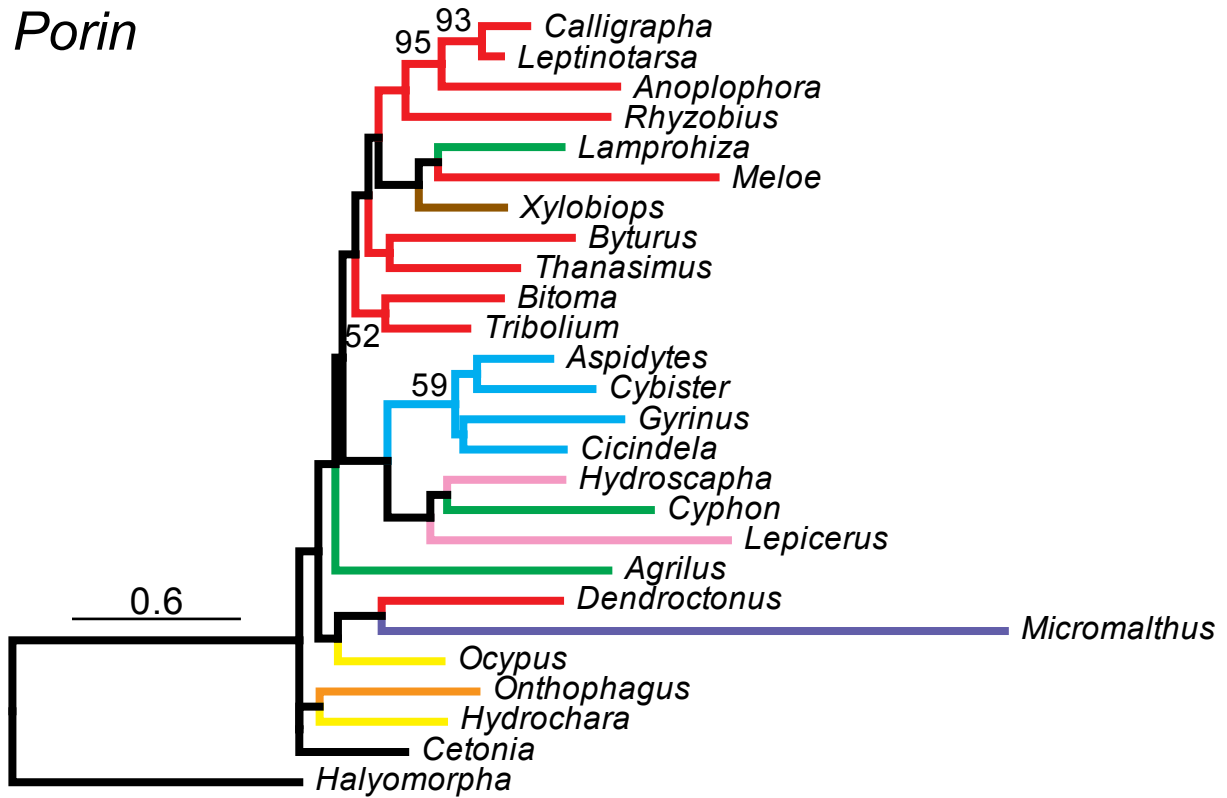
Pen



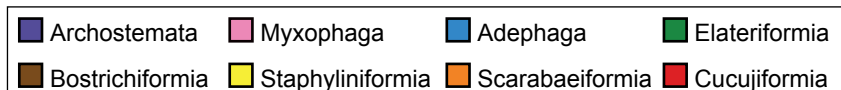
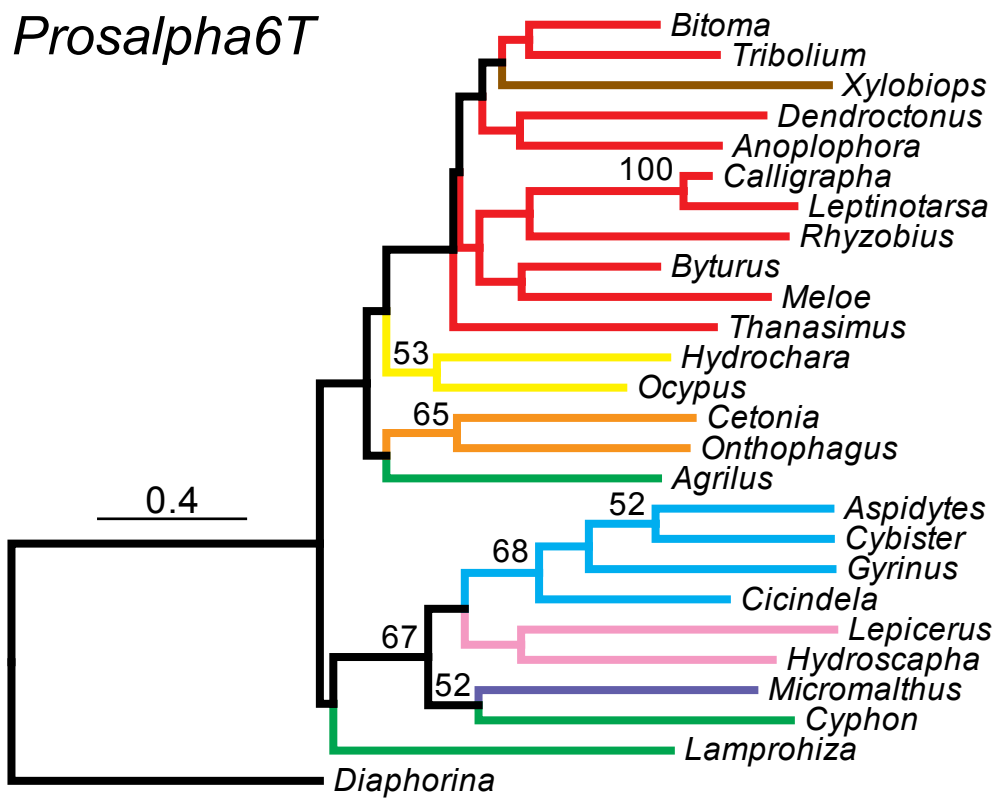
poe



Porin

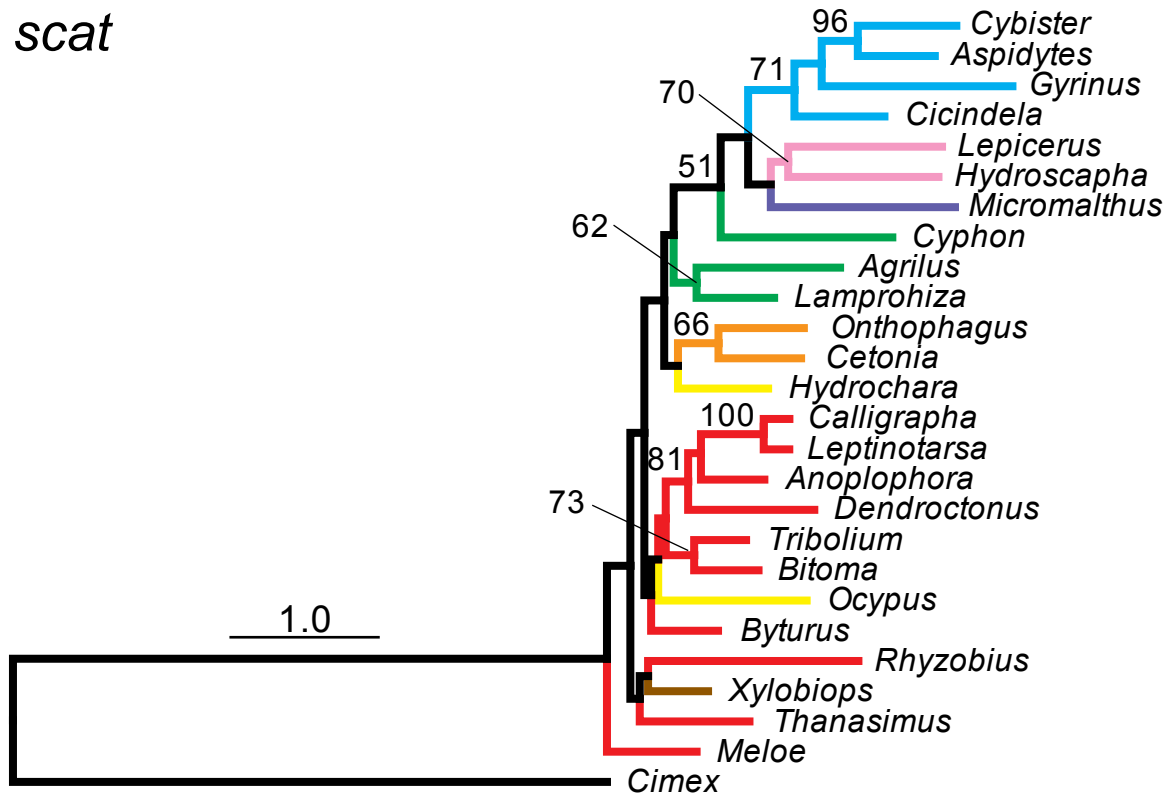


Prosalpha6T

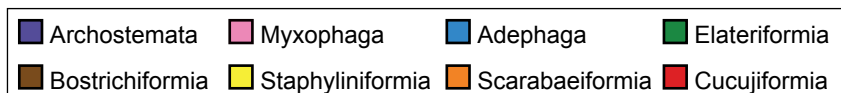
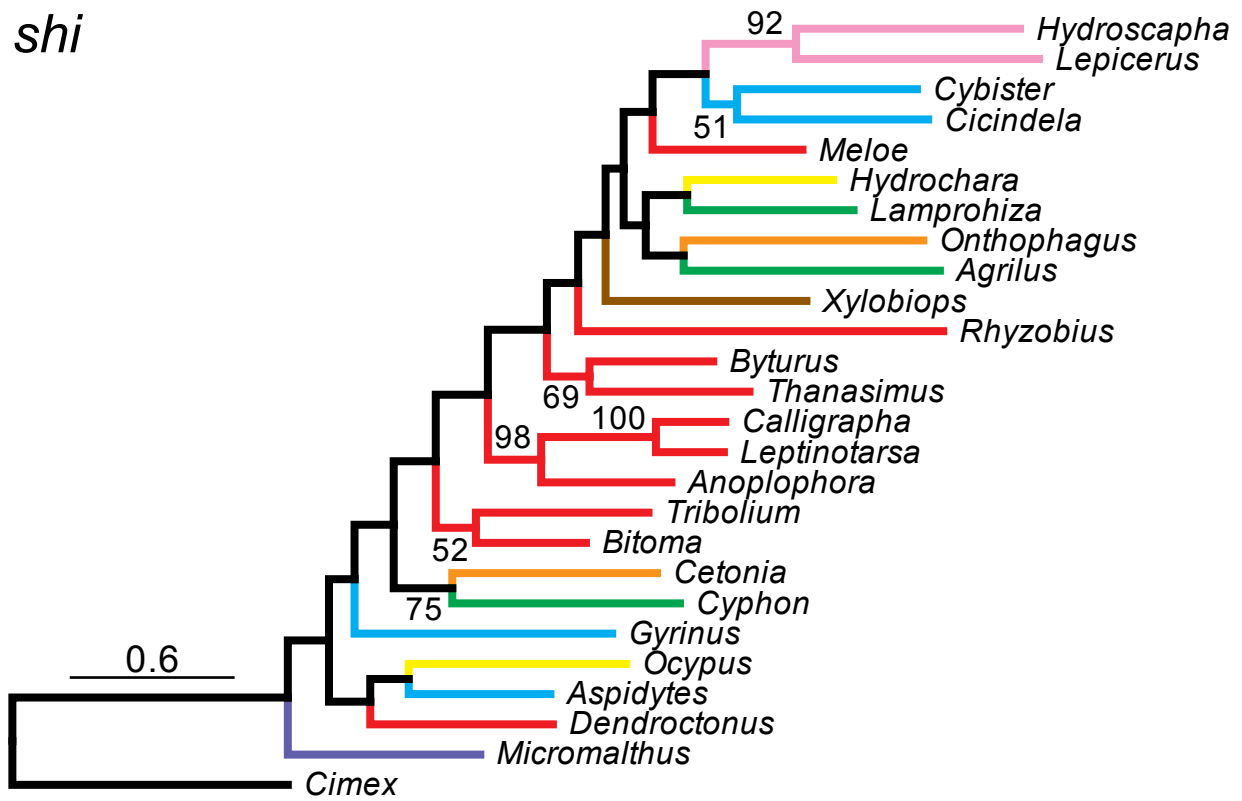


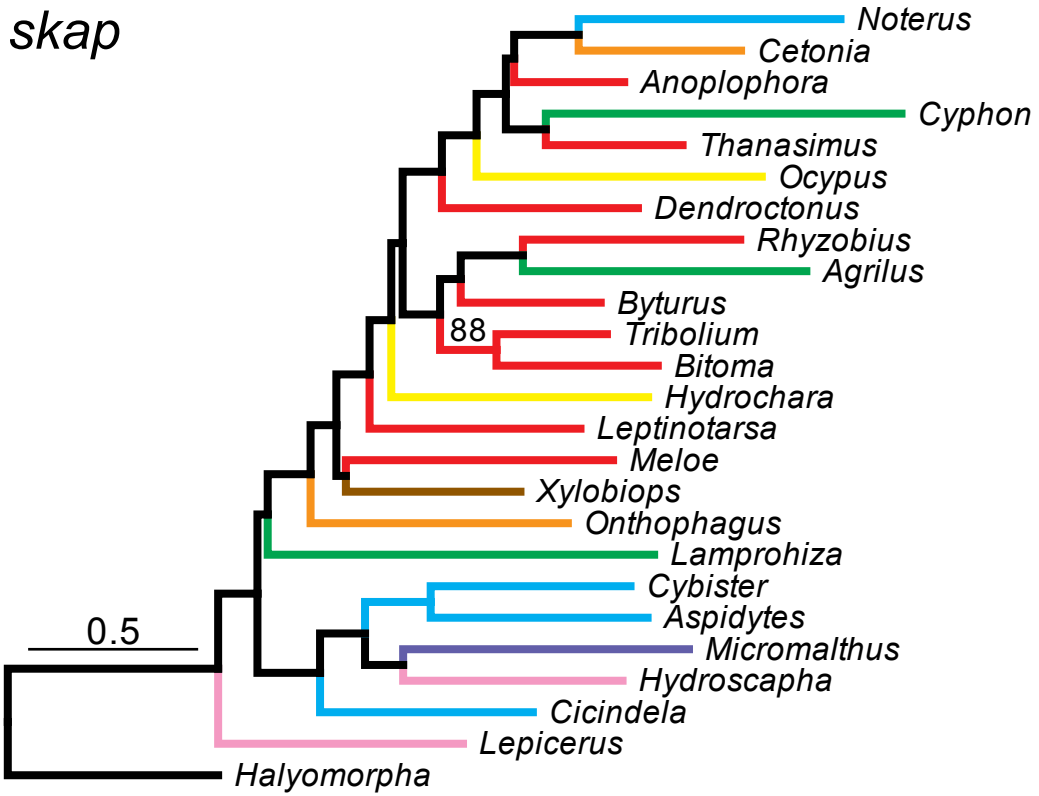


scat

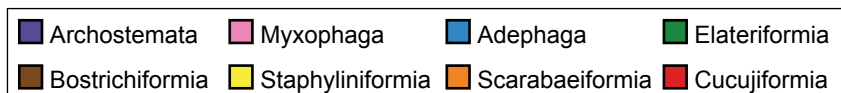
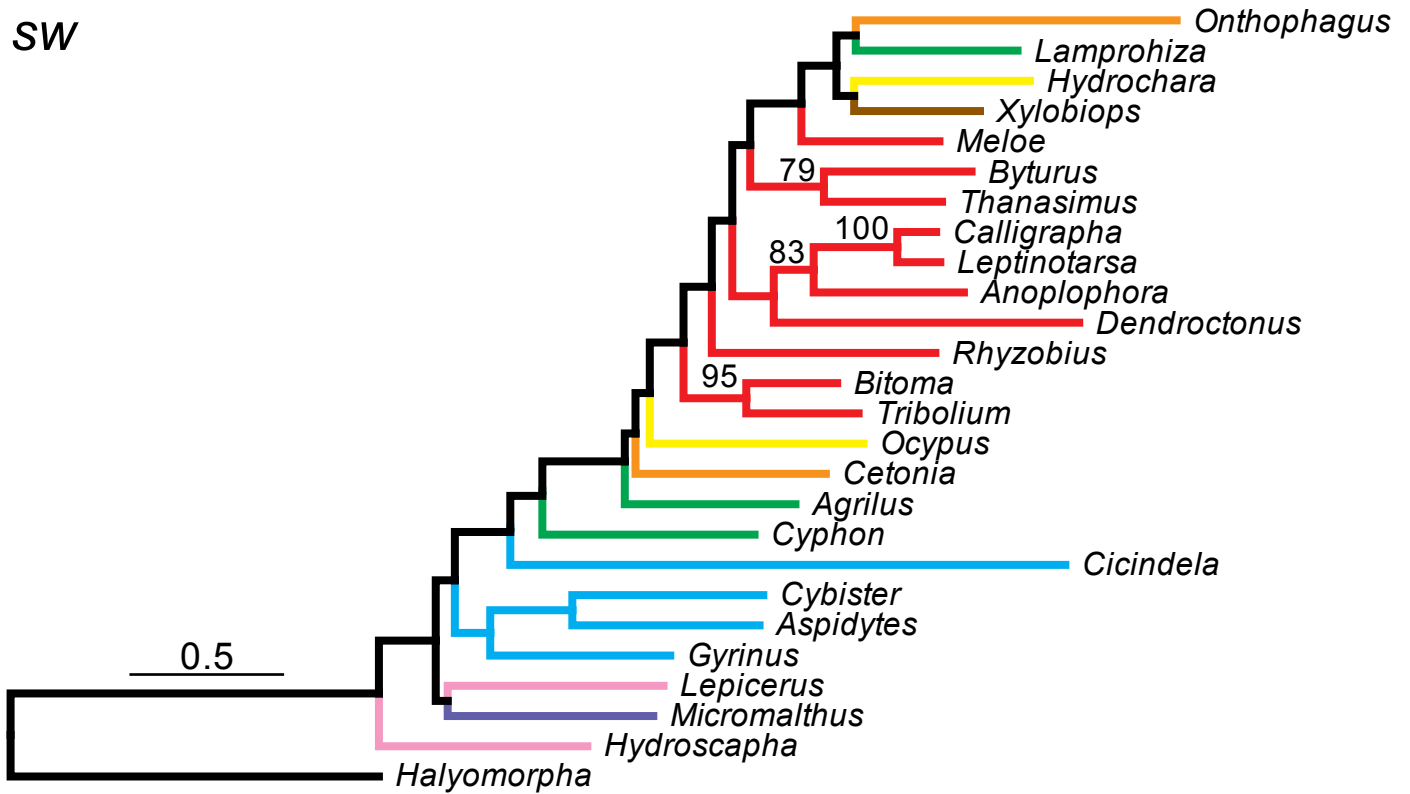


shi

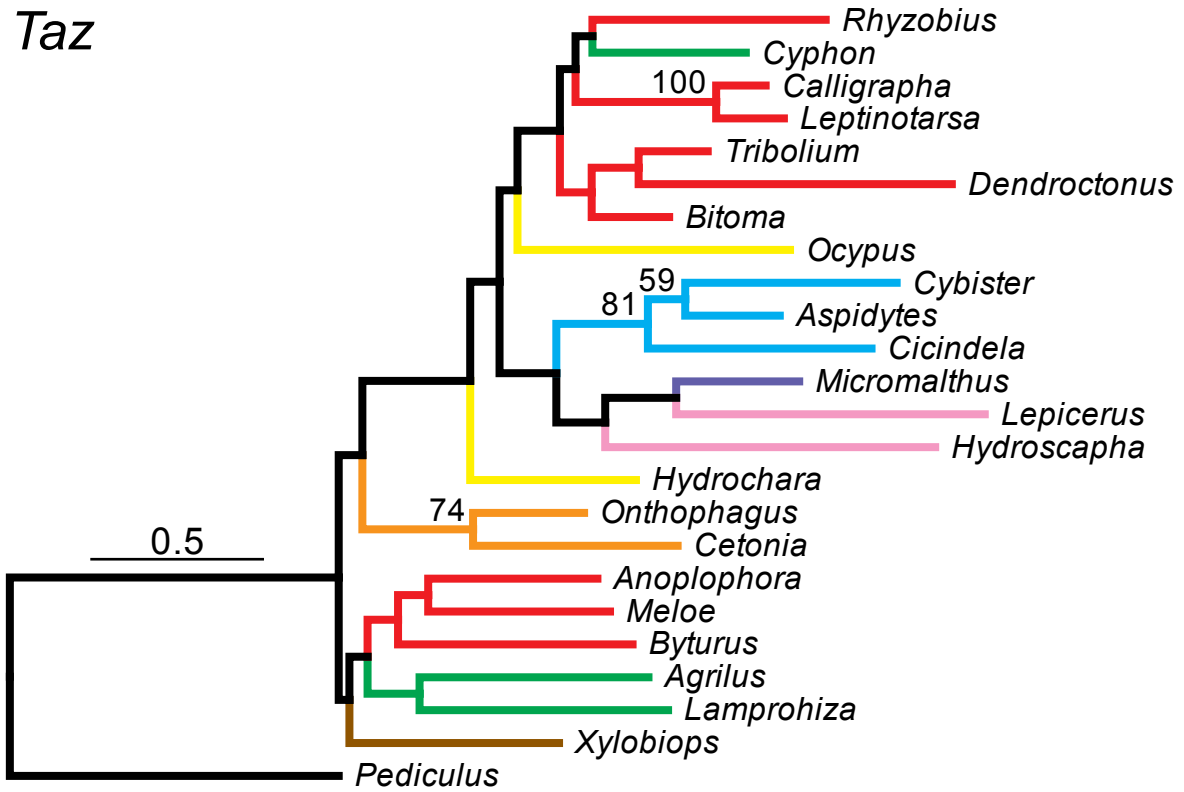




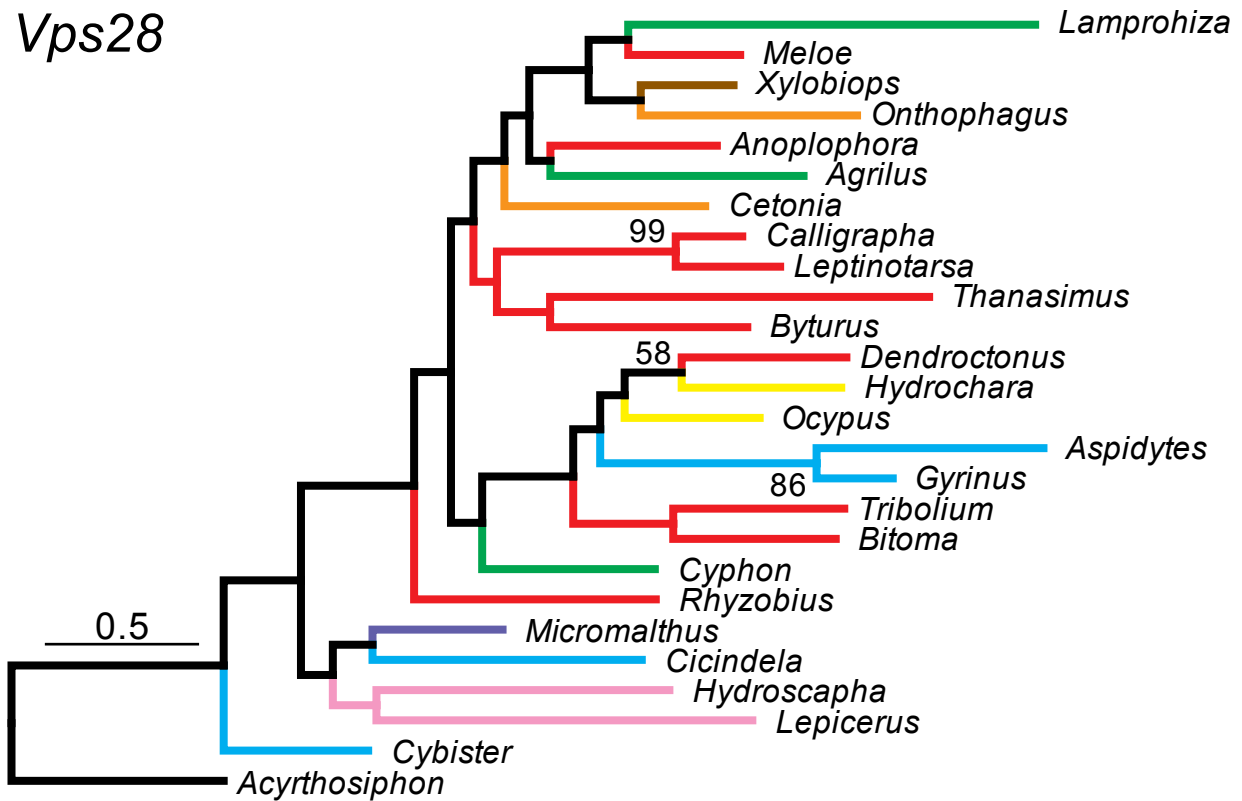
SW



Taz

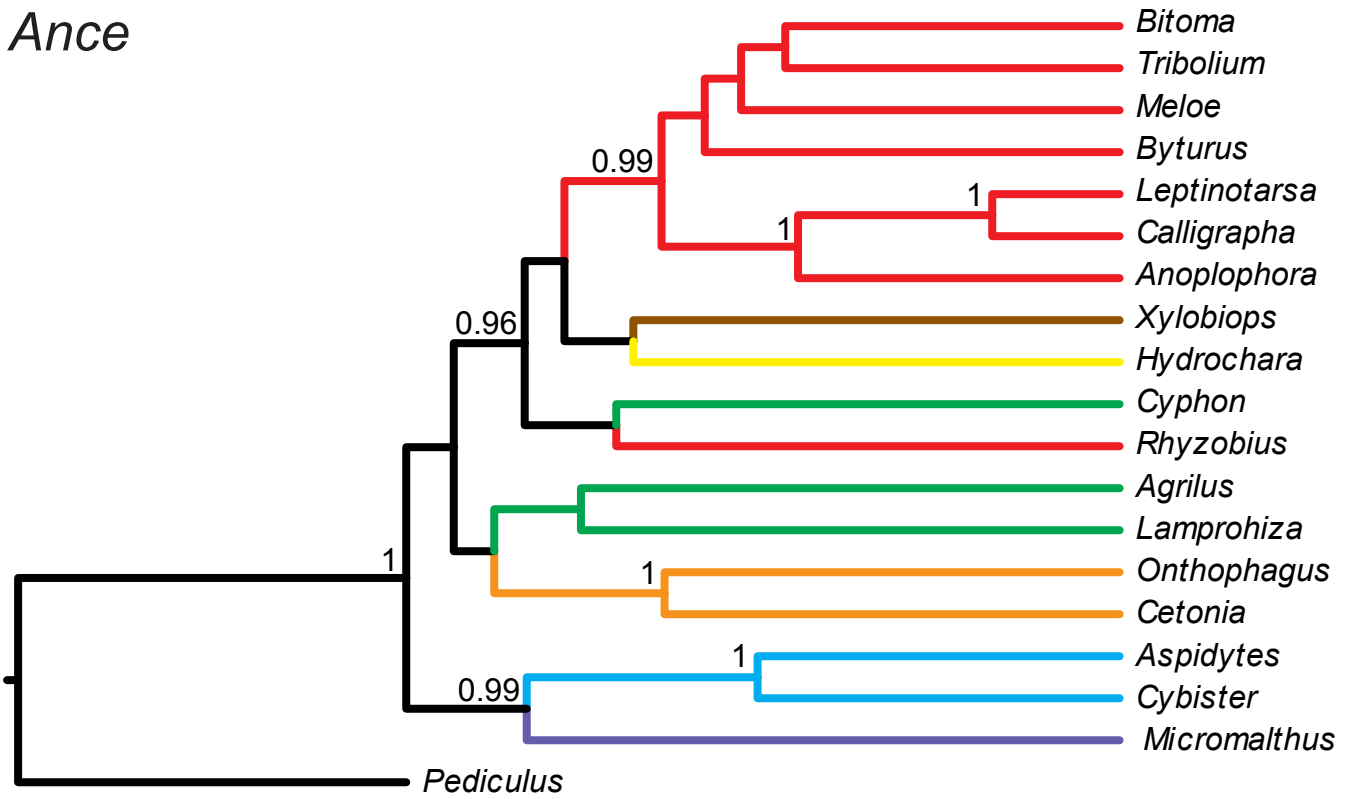


Vps28

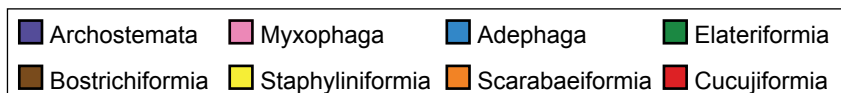
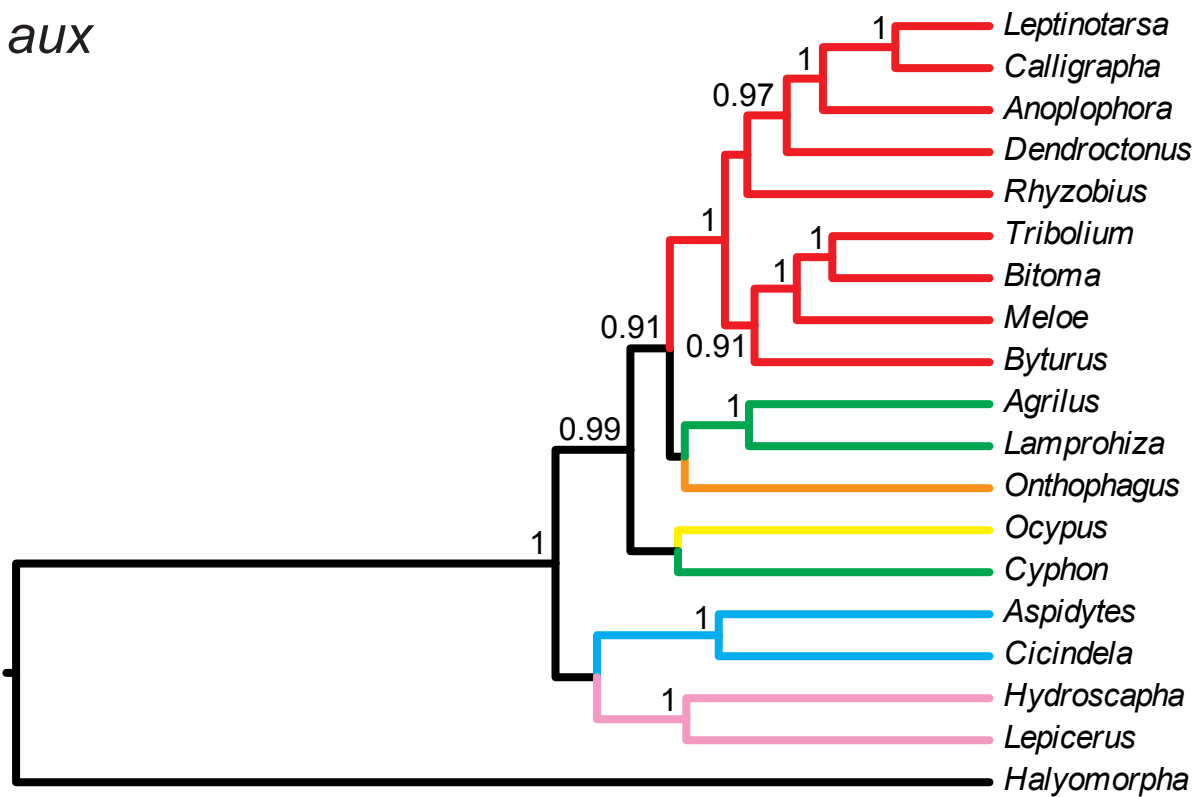


**File S4.** Bayesian inference trees based on the nucleotide alignments of different sperm individualization genes in beetles.

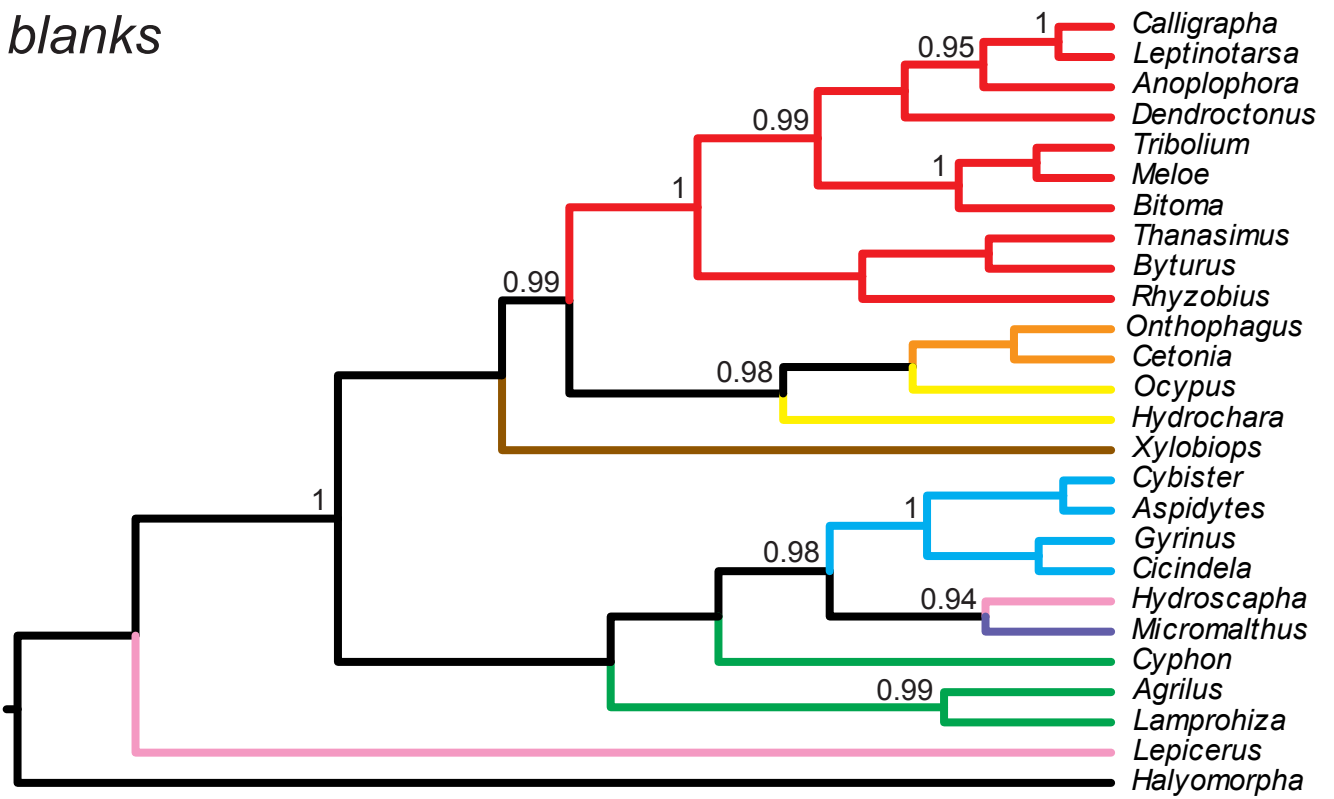
Ance



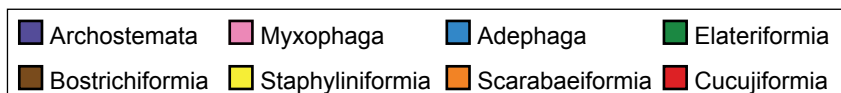
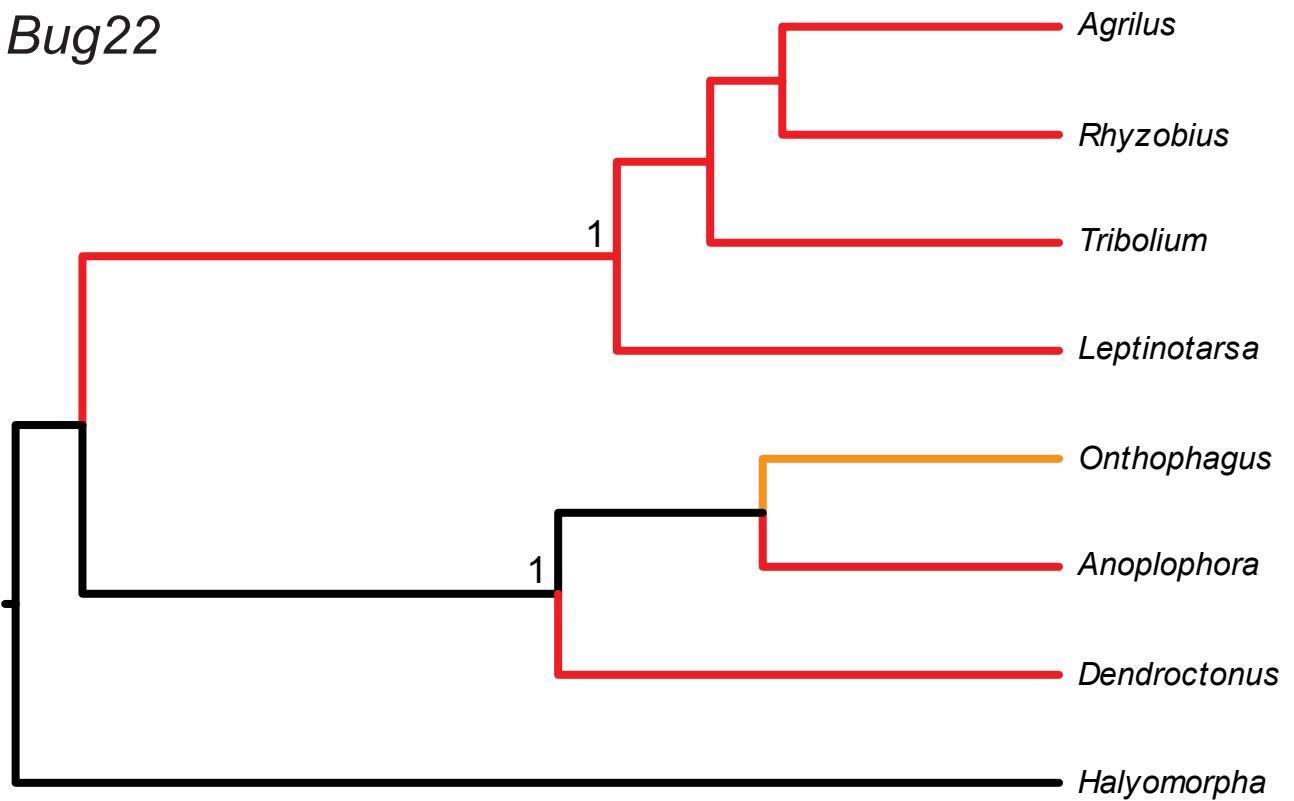
aux



blanks

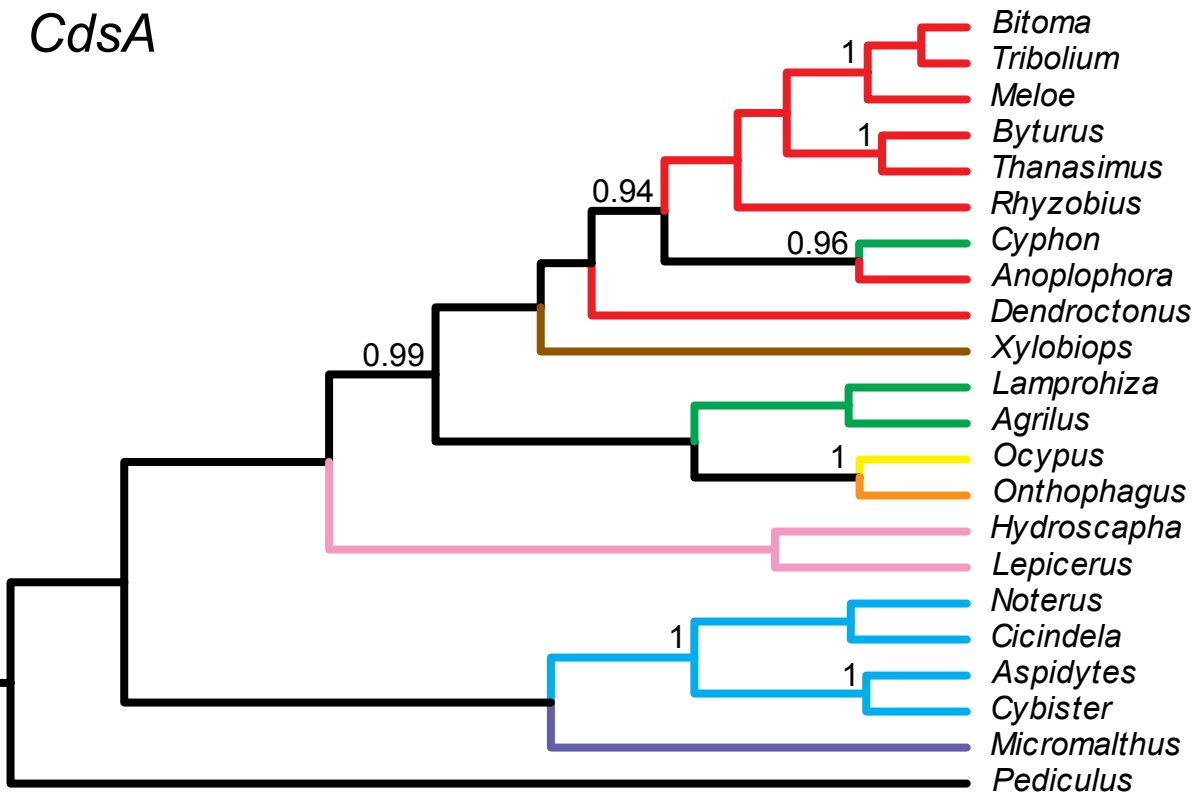


Bug22

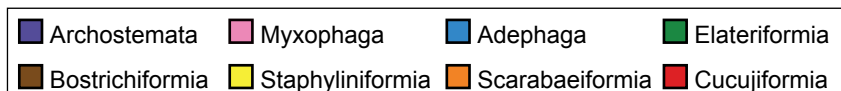
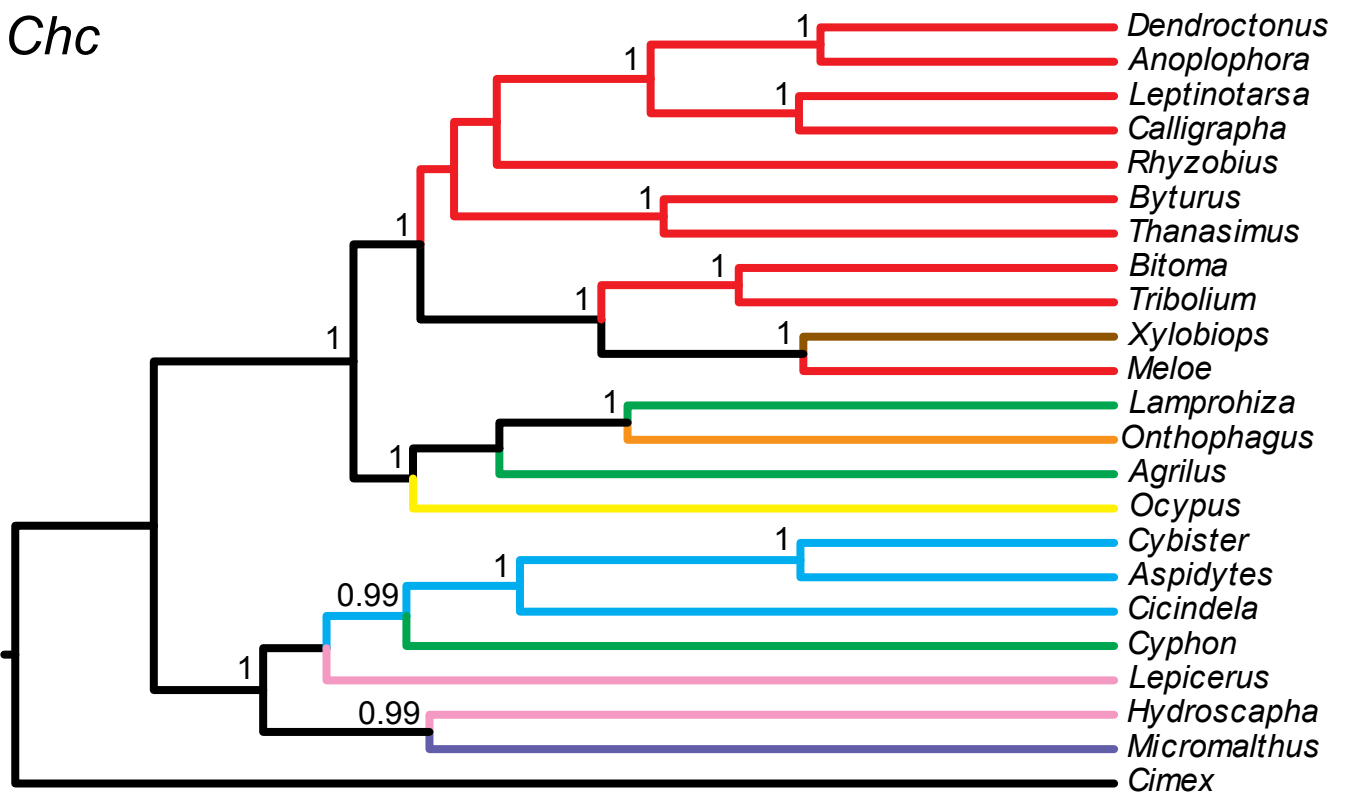




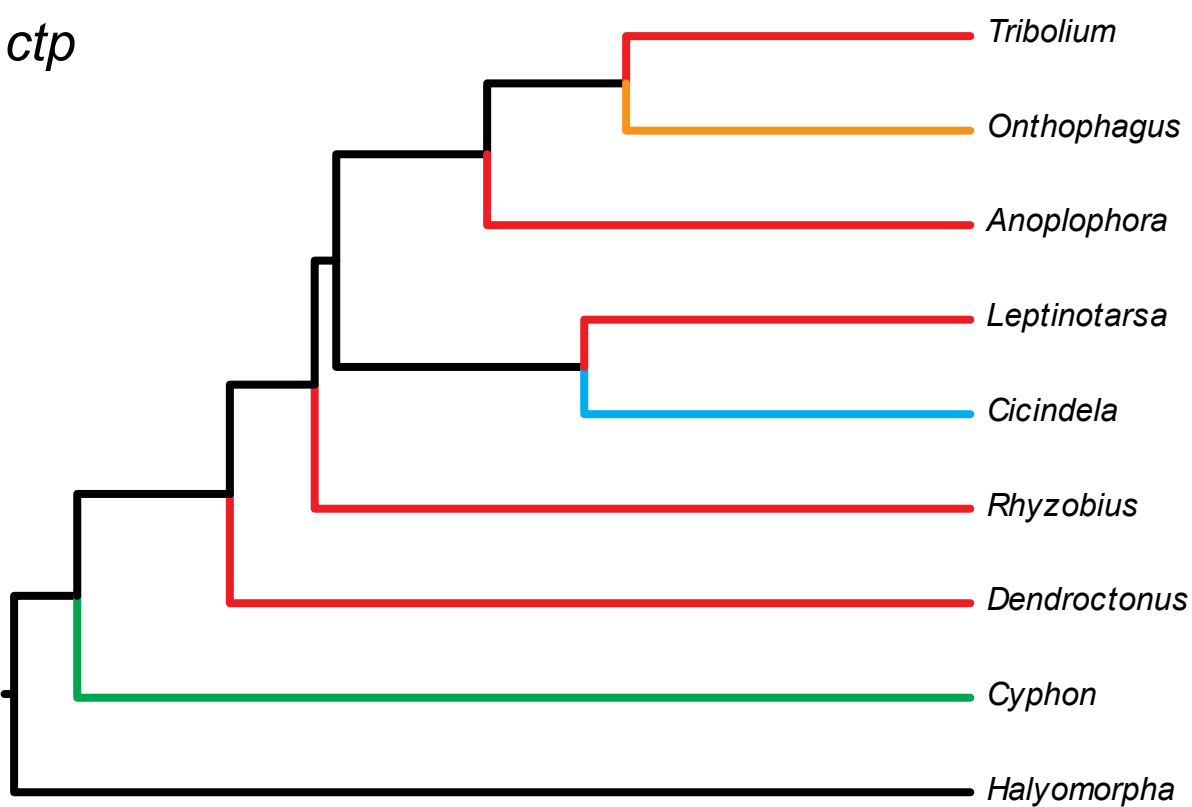
# CdsA



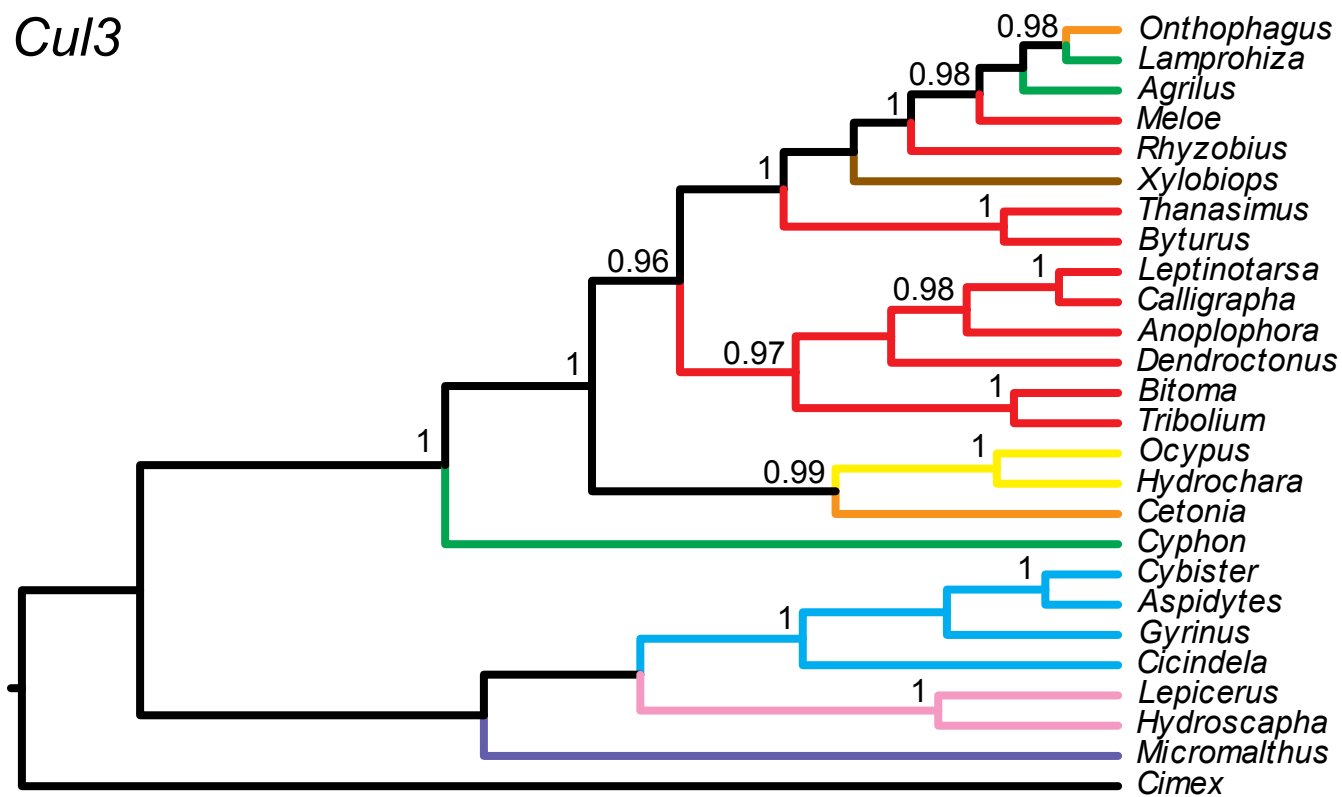
# Chc



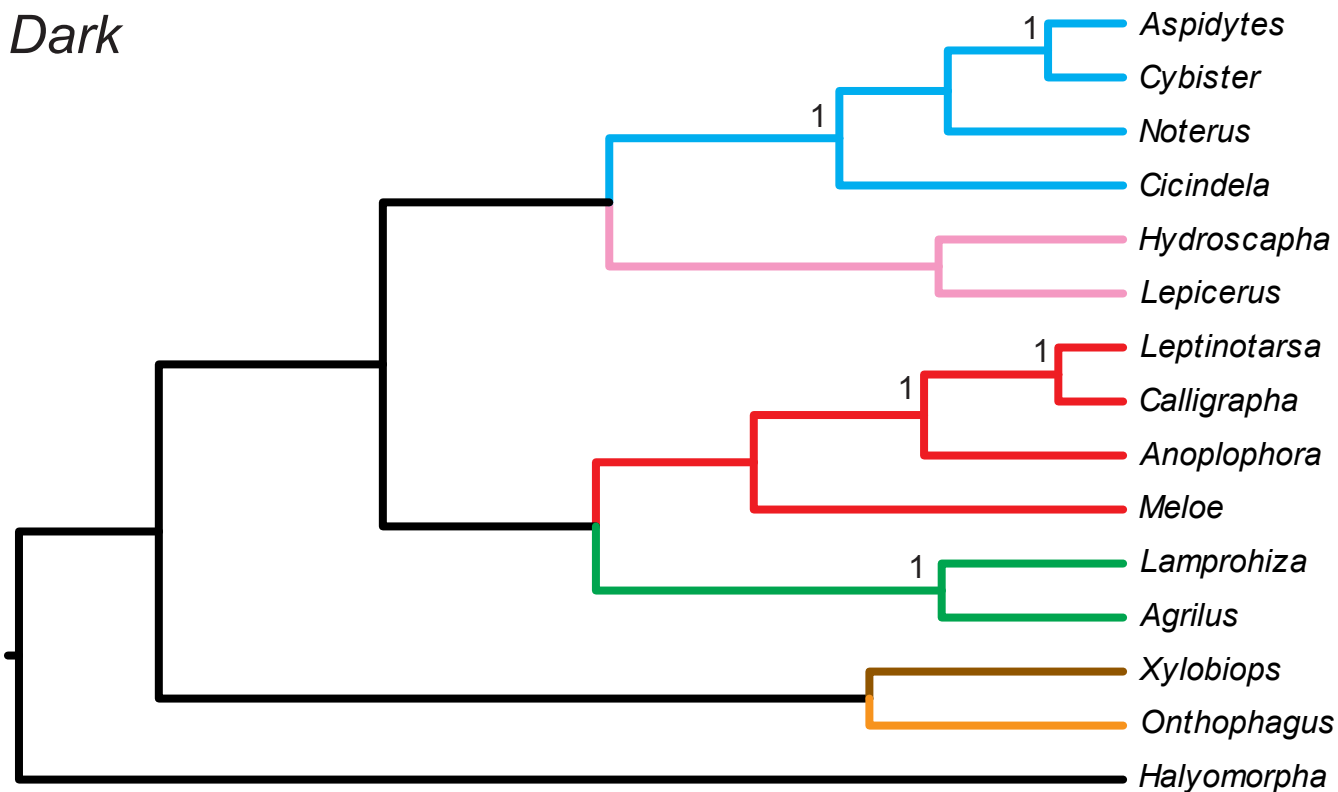
*ctp*



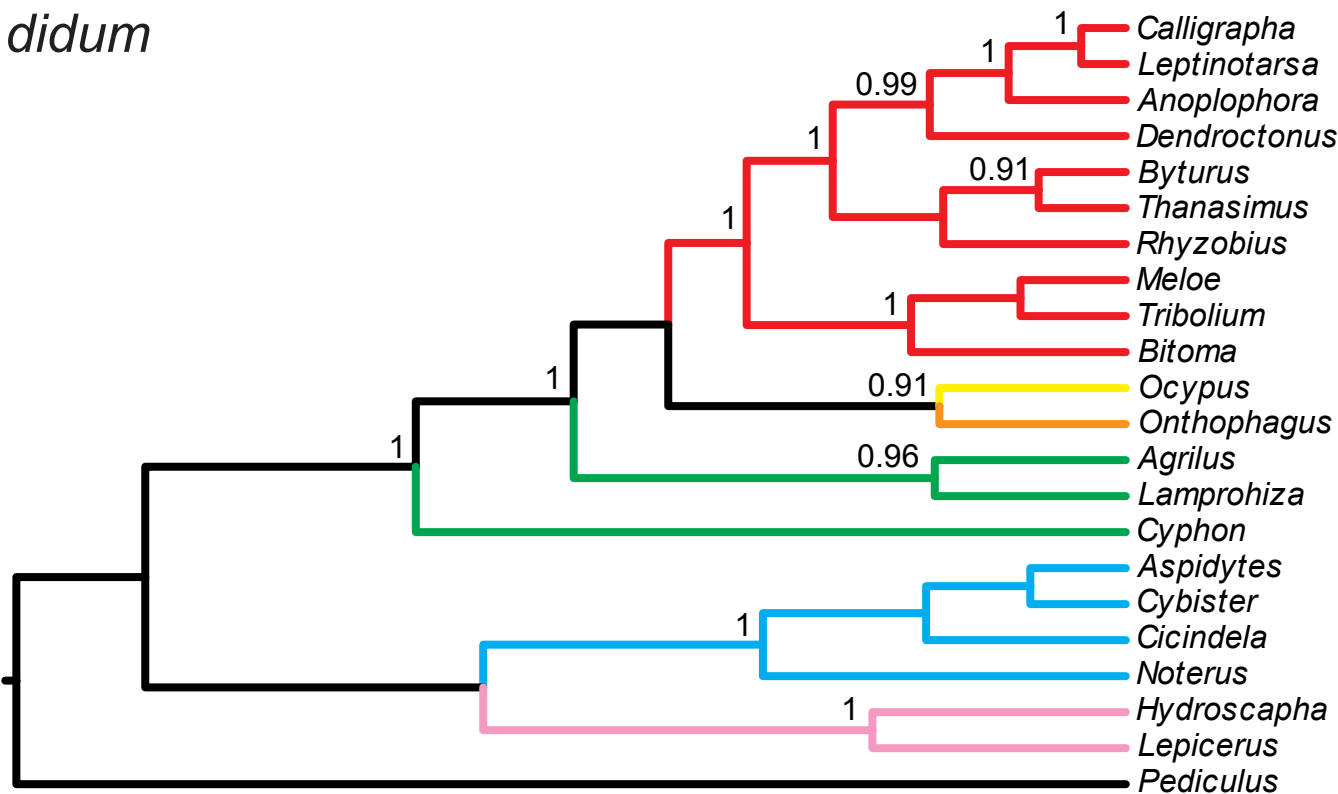
*Cul3*



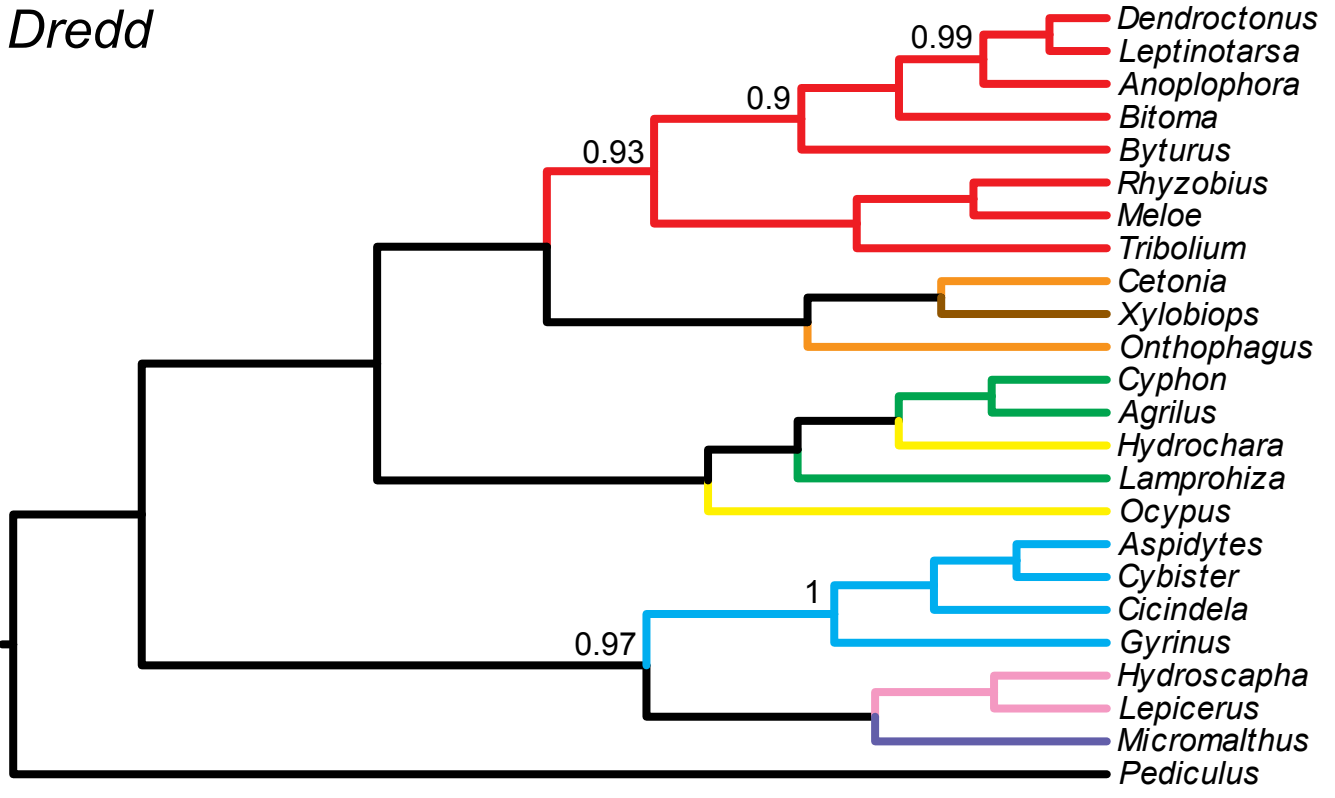
Dark



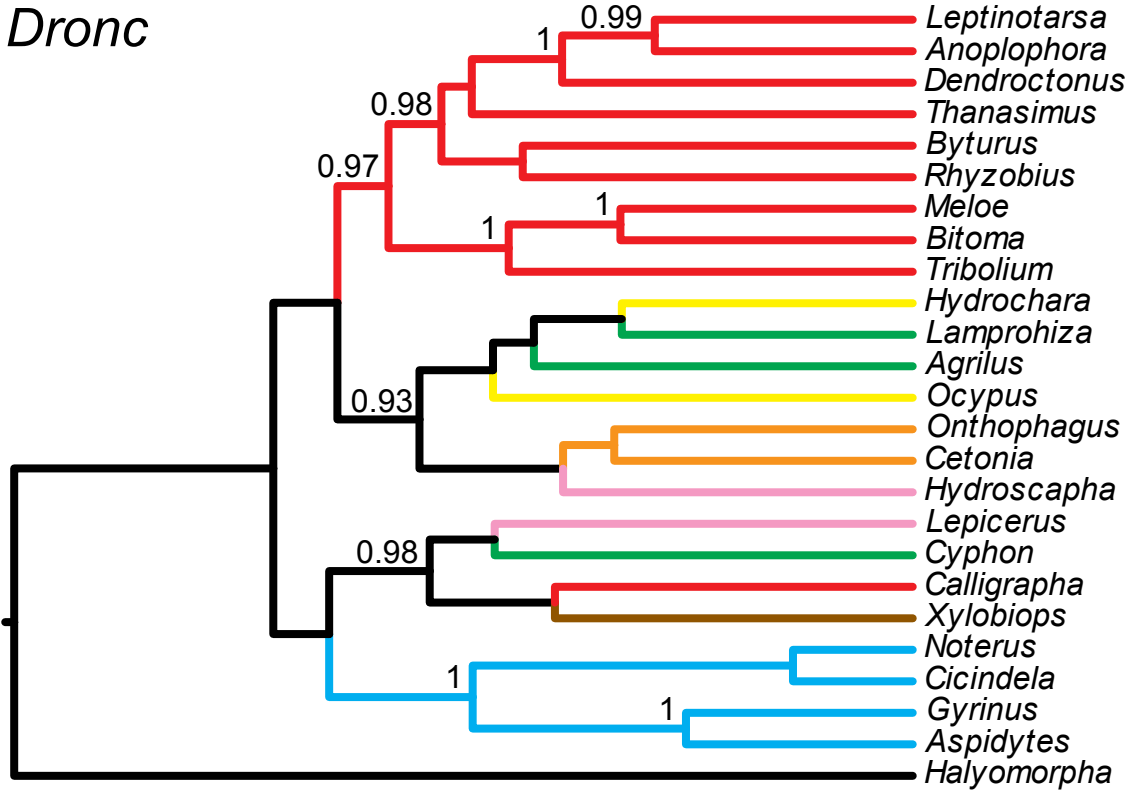
didum



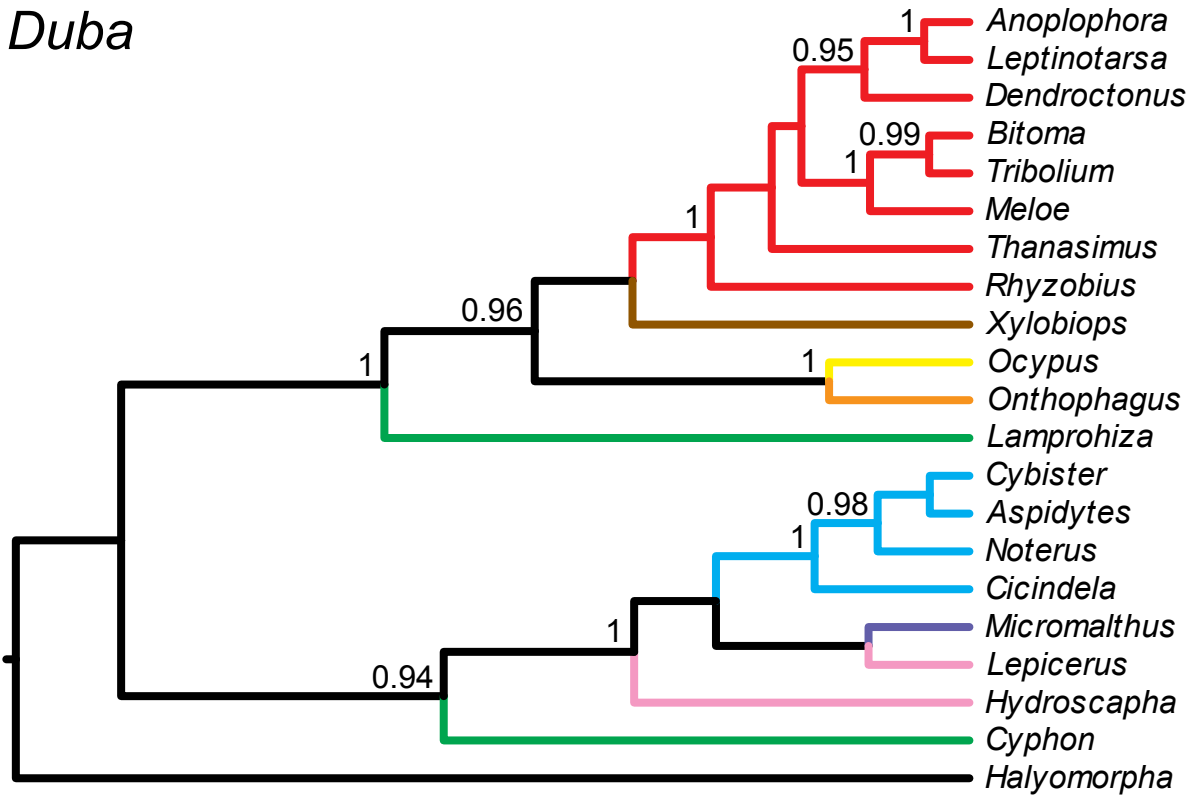
# Dredd



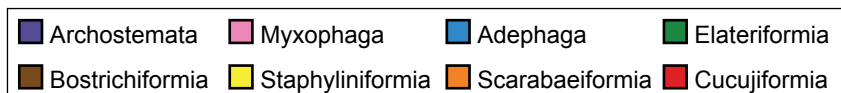
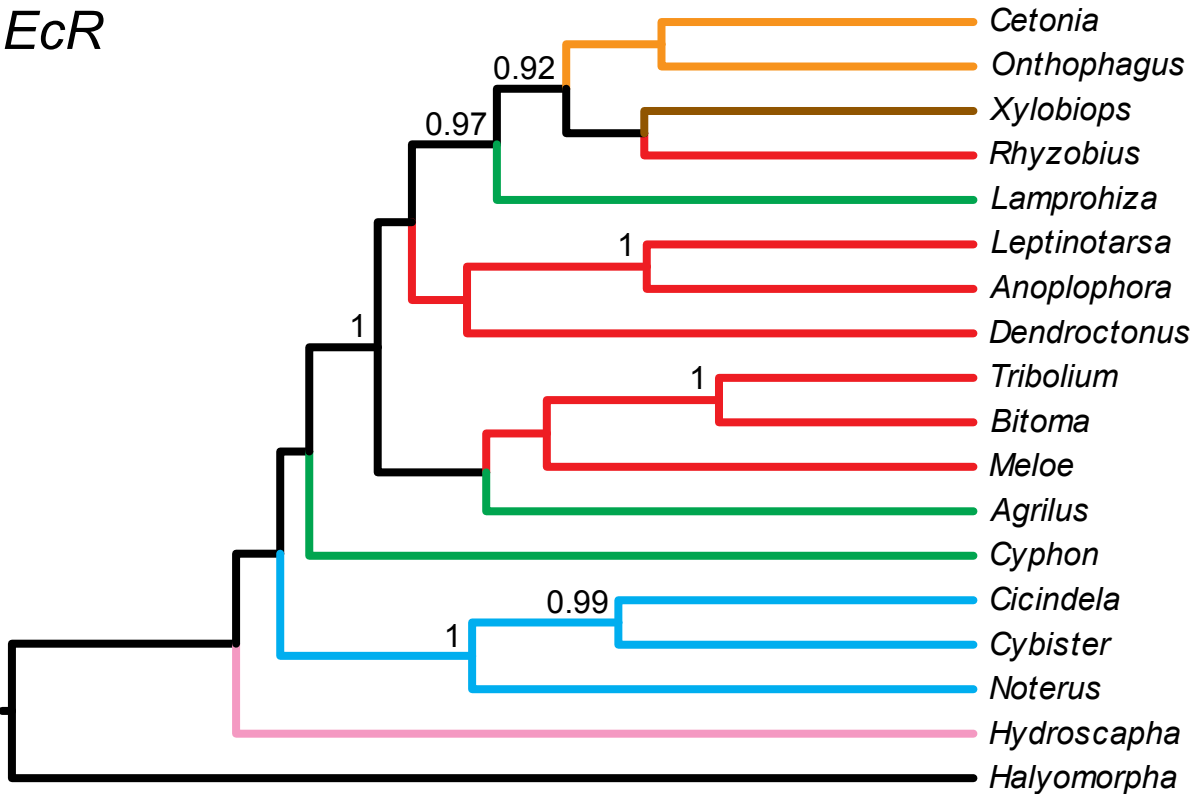
# Dronc



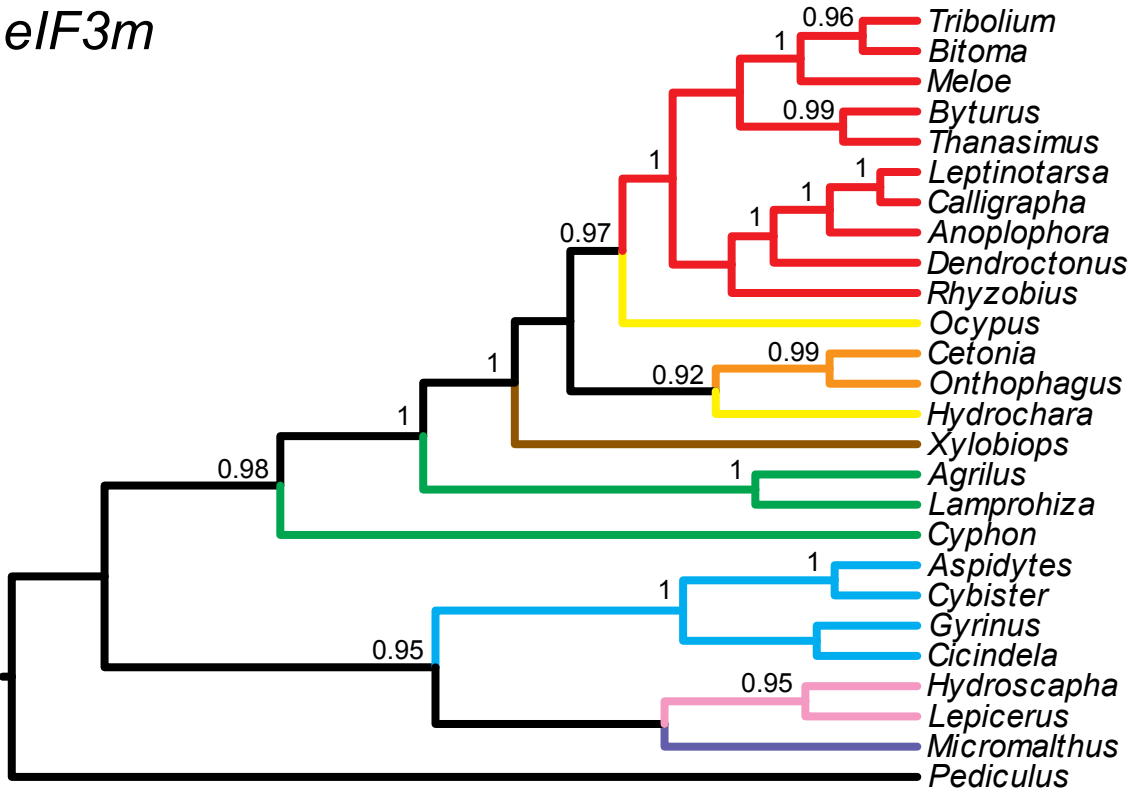
*Duba*



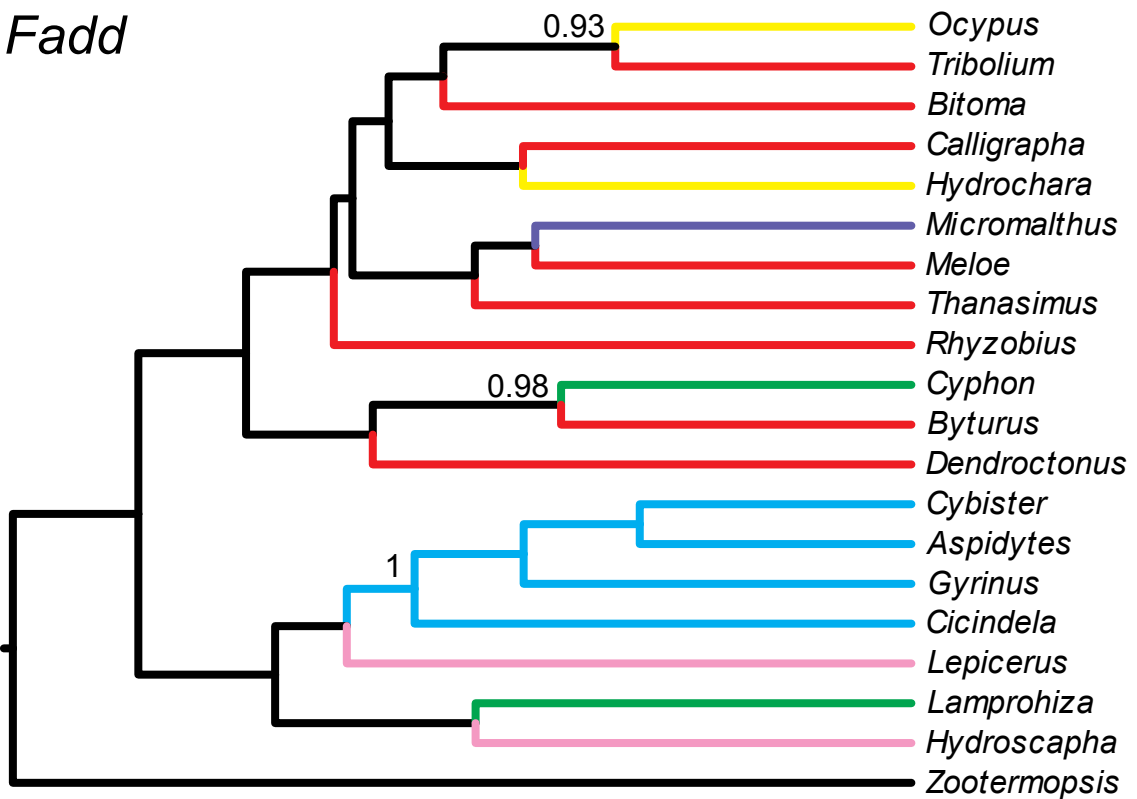
*EcR*



*eIF3m*

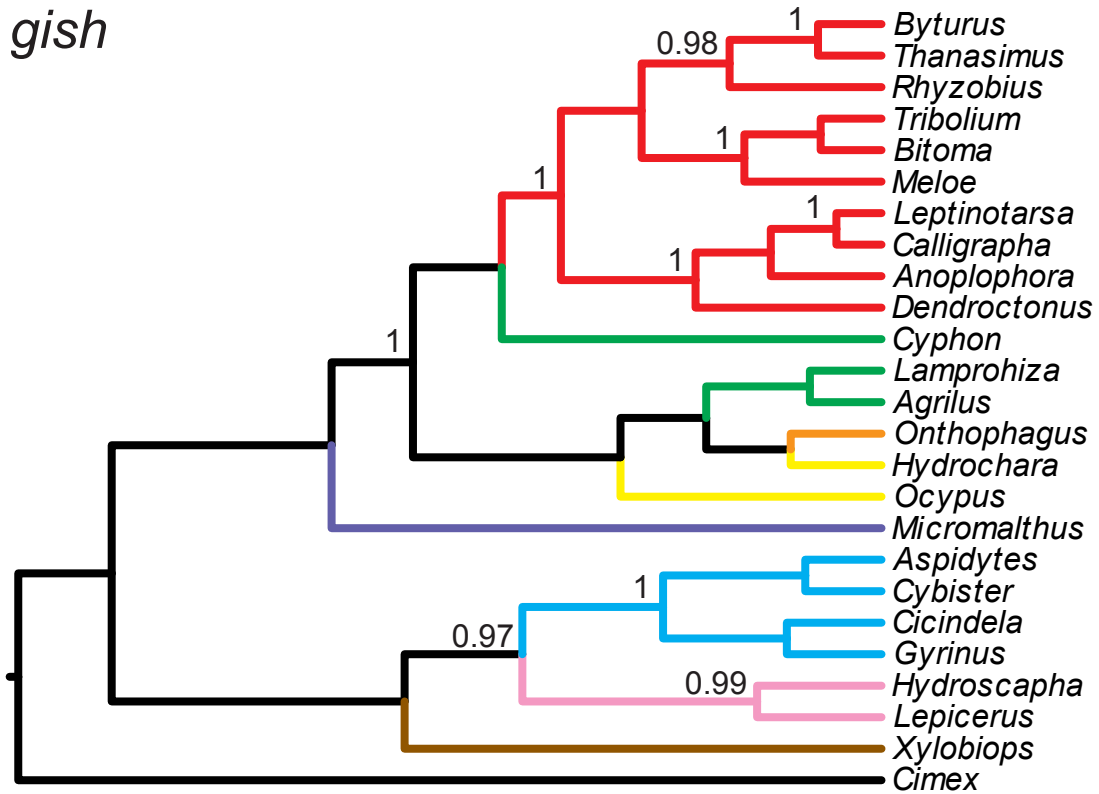


*Fadd*

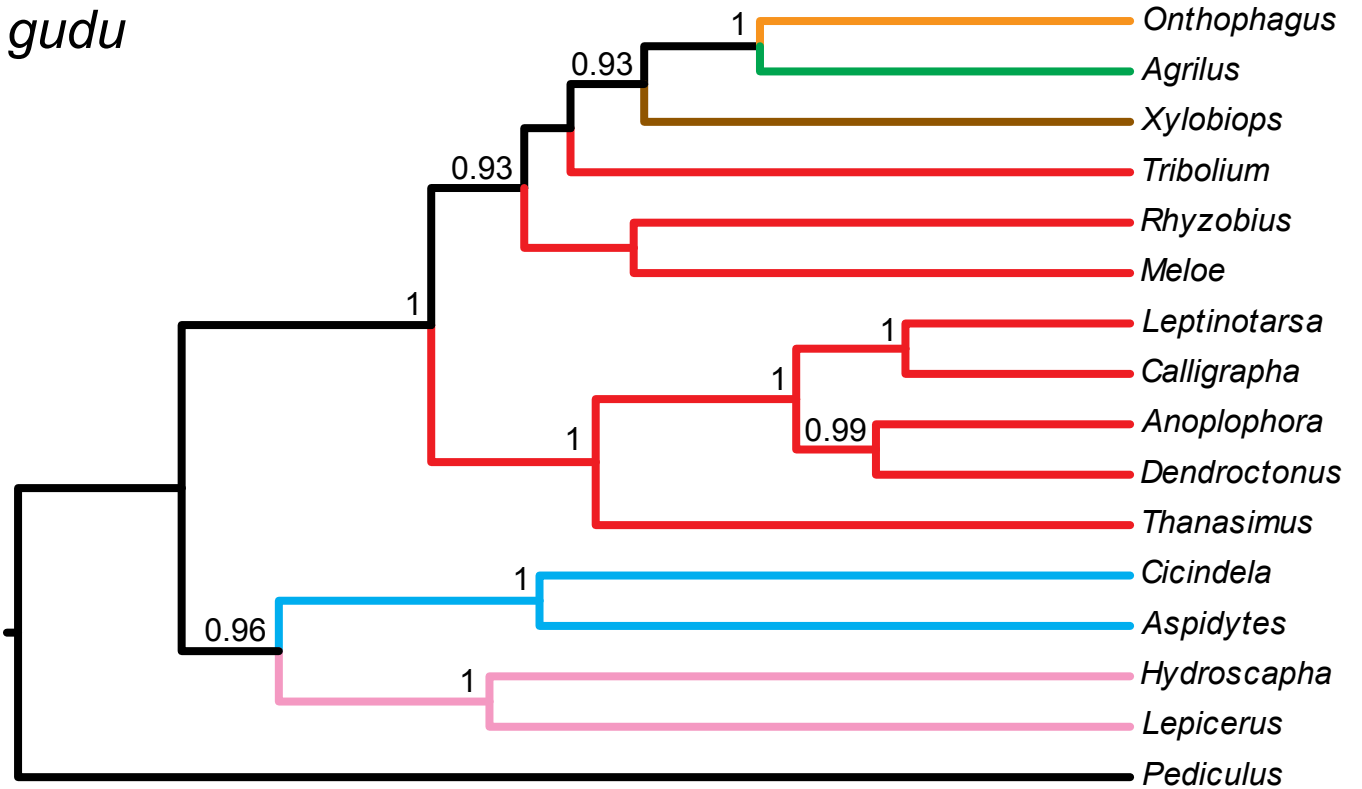




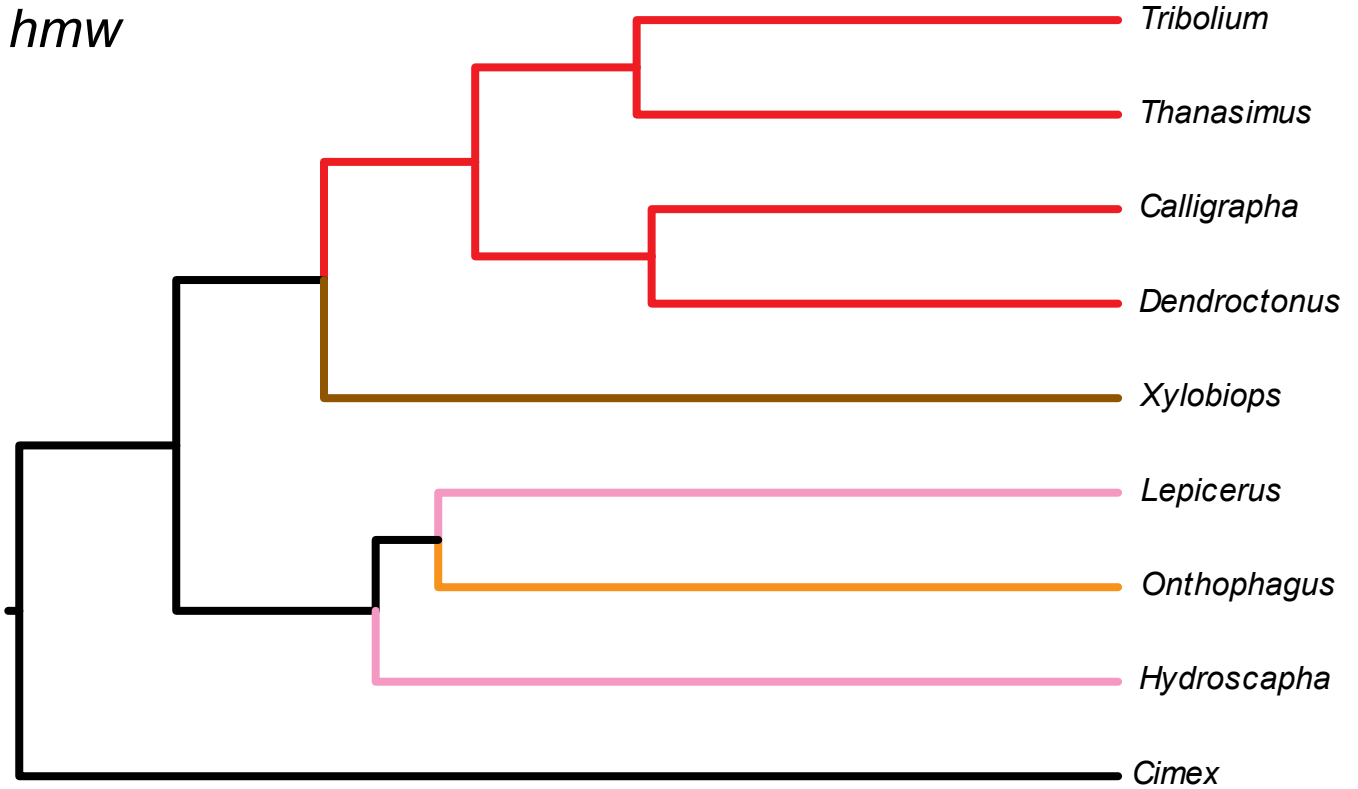
*gish*



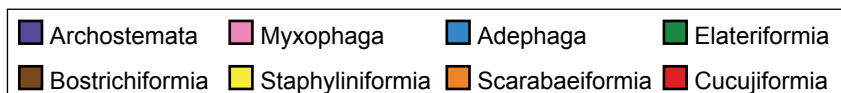
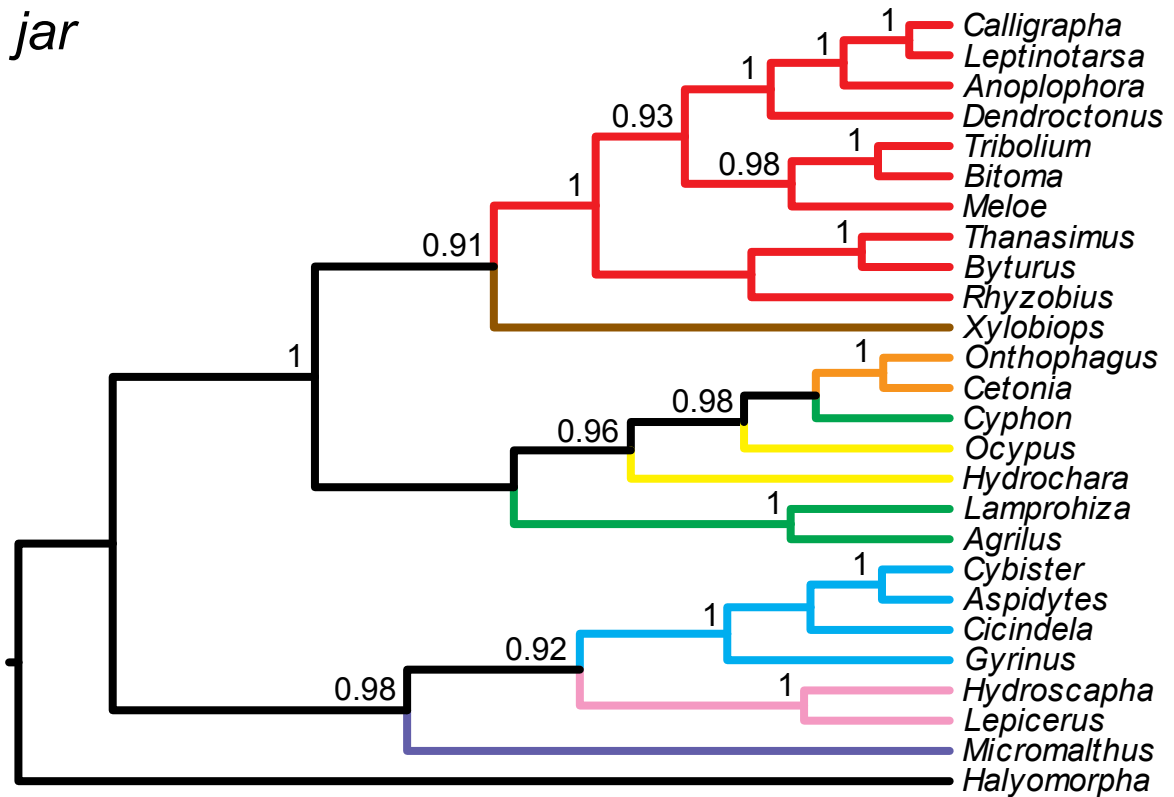
*gudu*



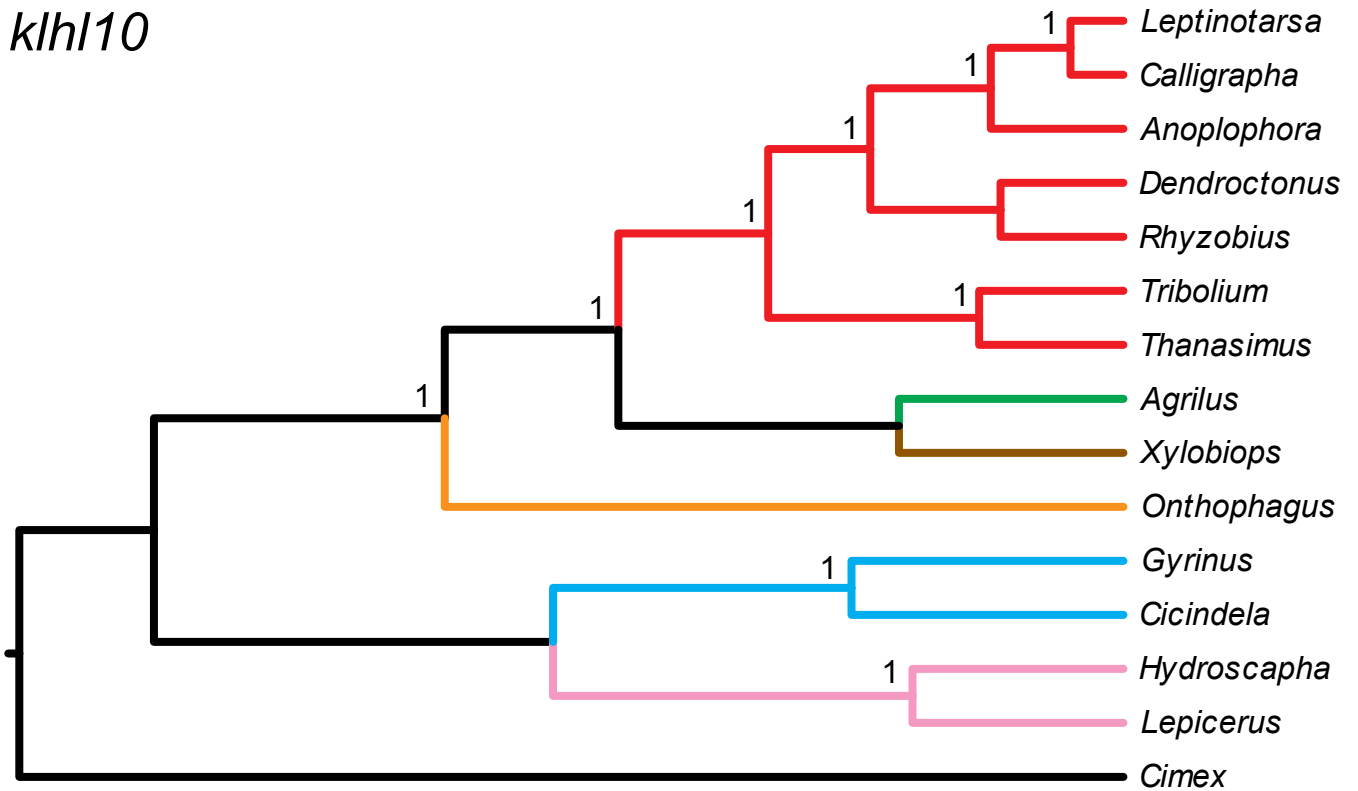
hmw



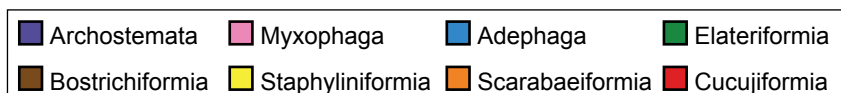
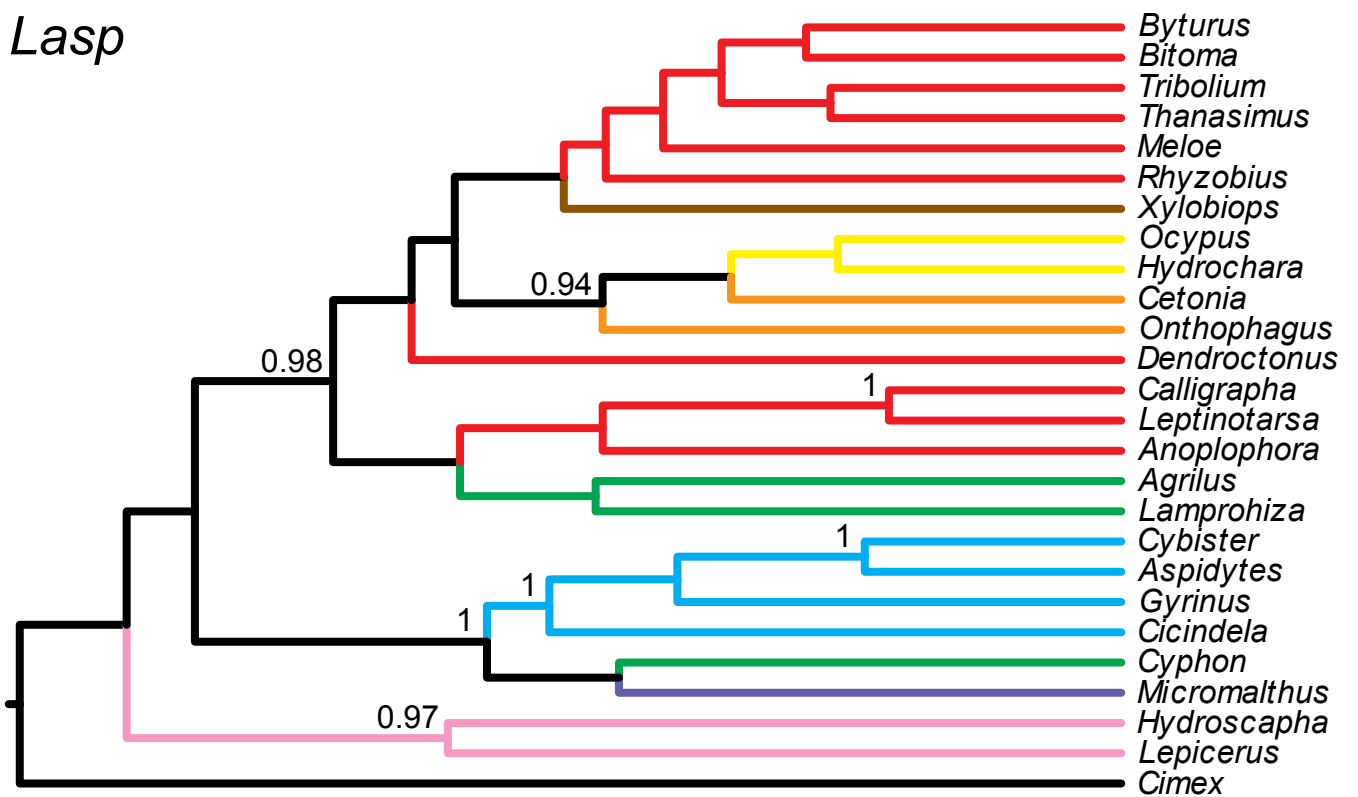
jar



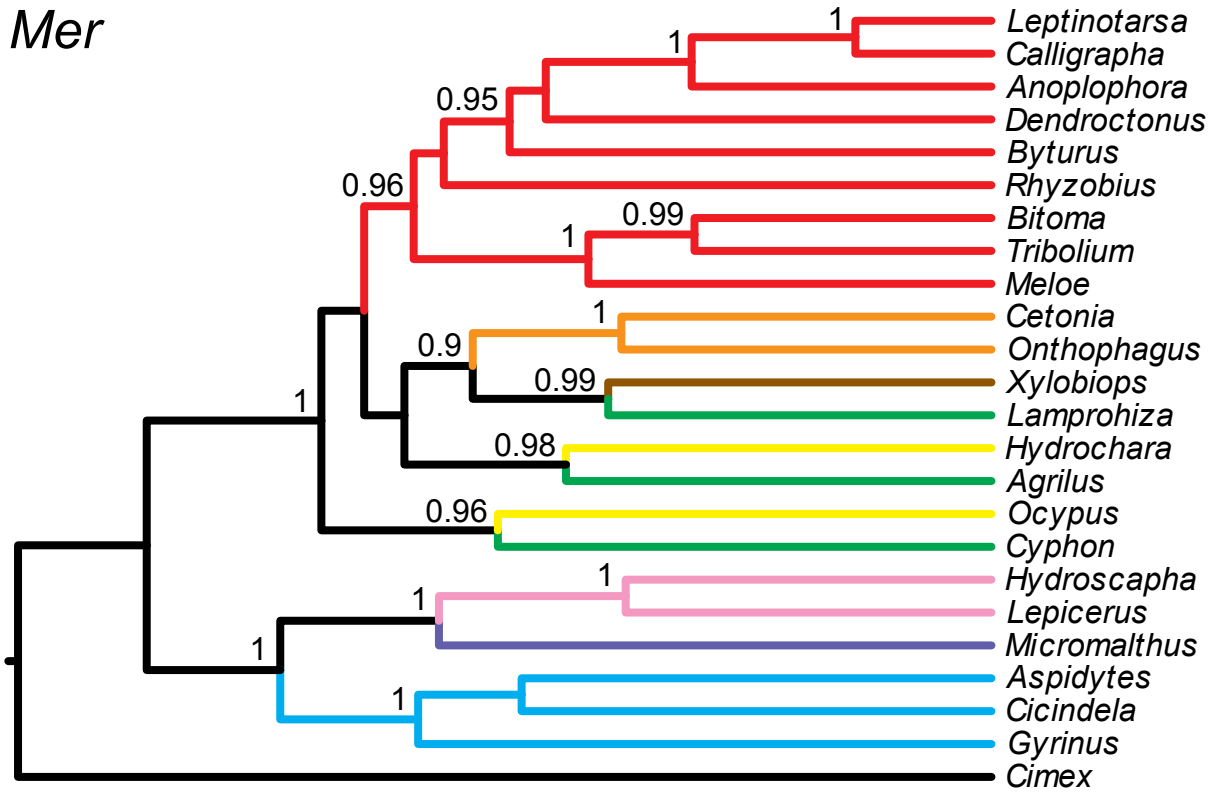
*klhl10*



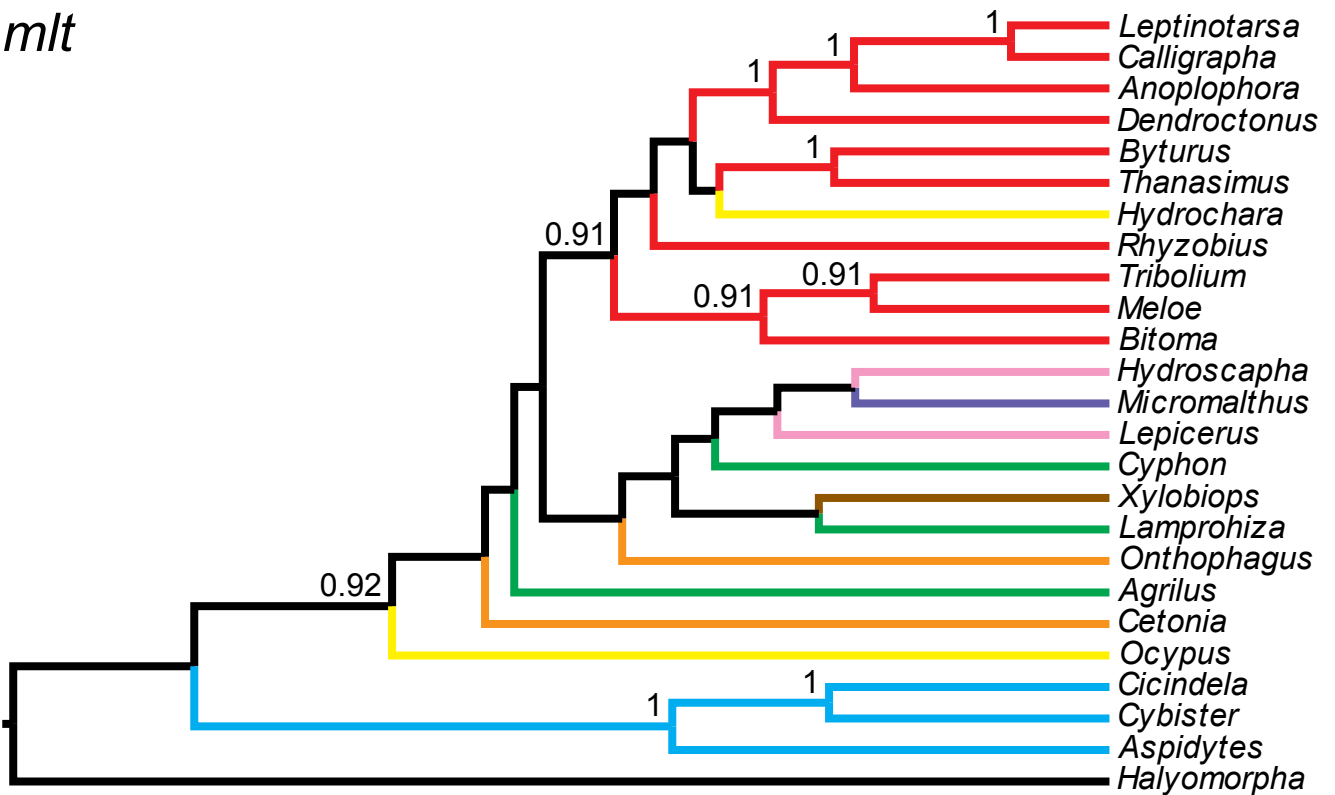
*Lasp*



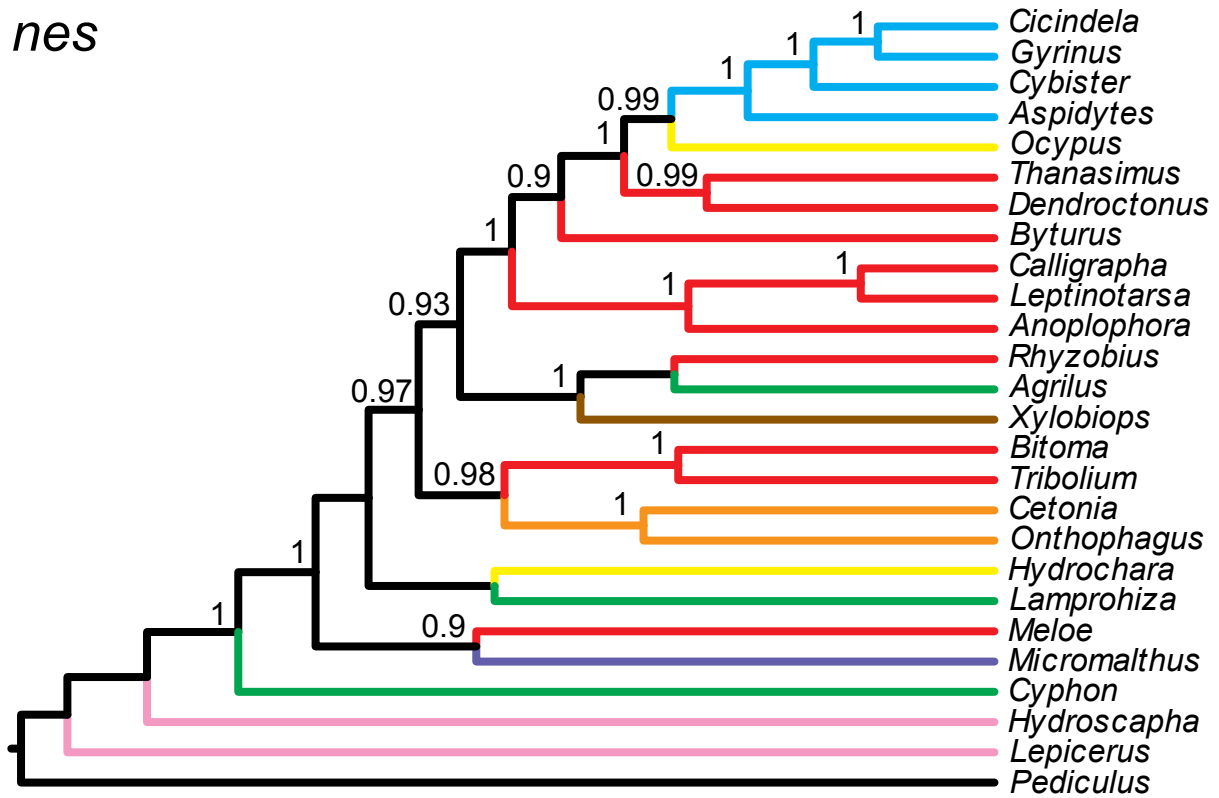
*Mer*



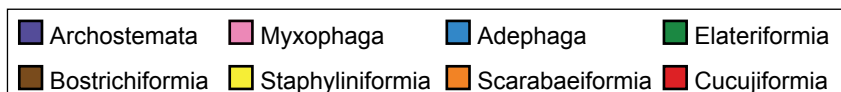
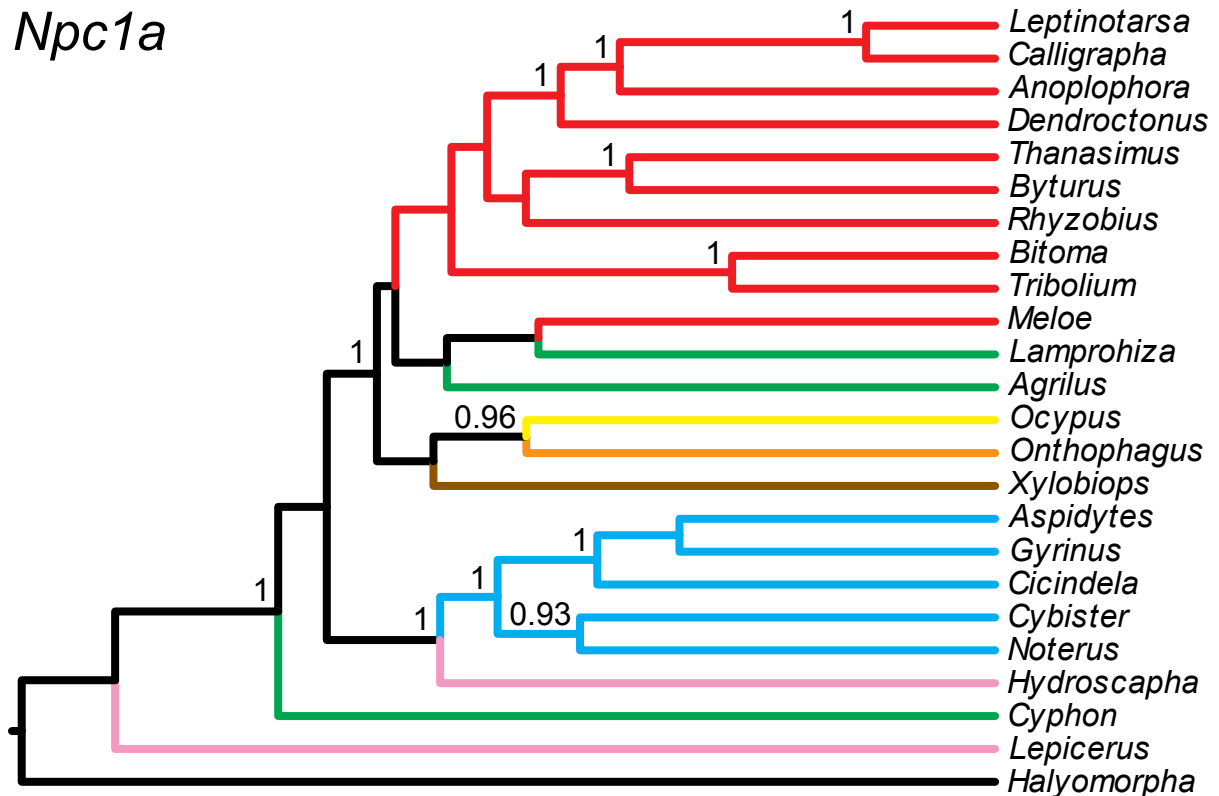
*mlt*



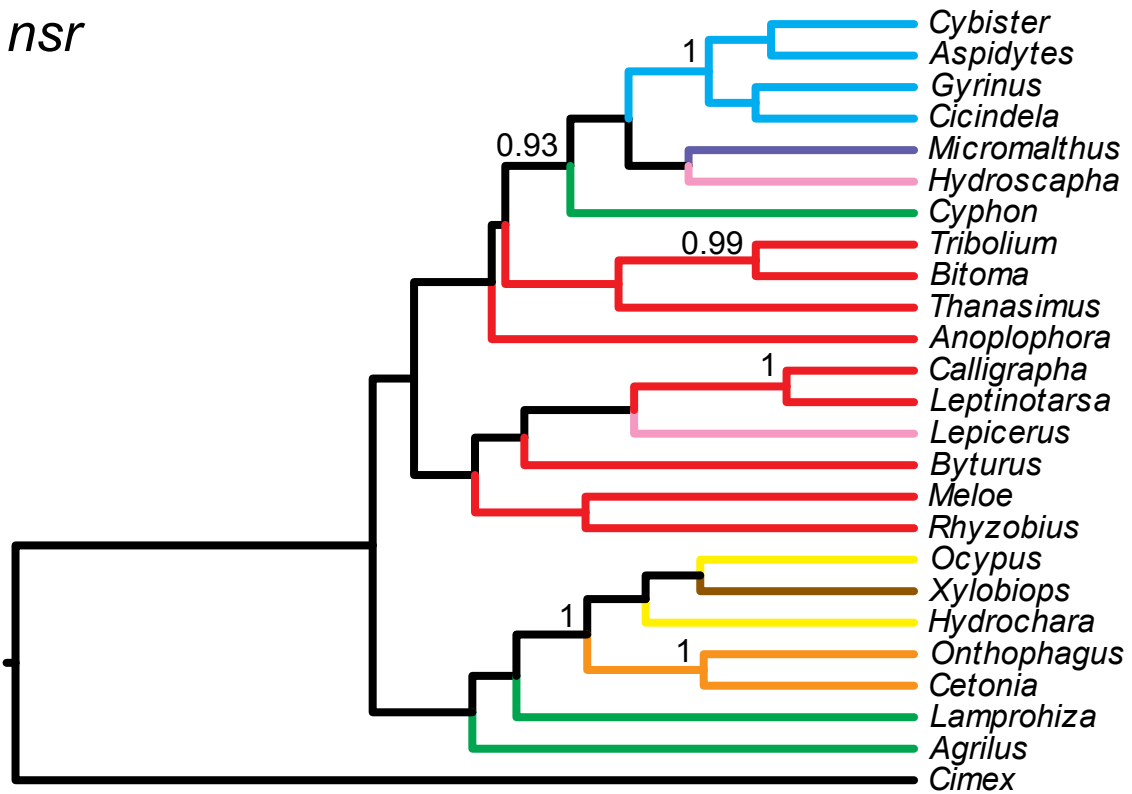
nes



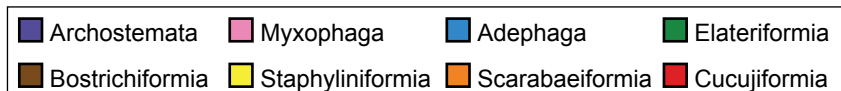
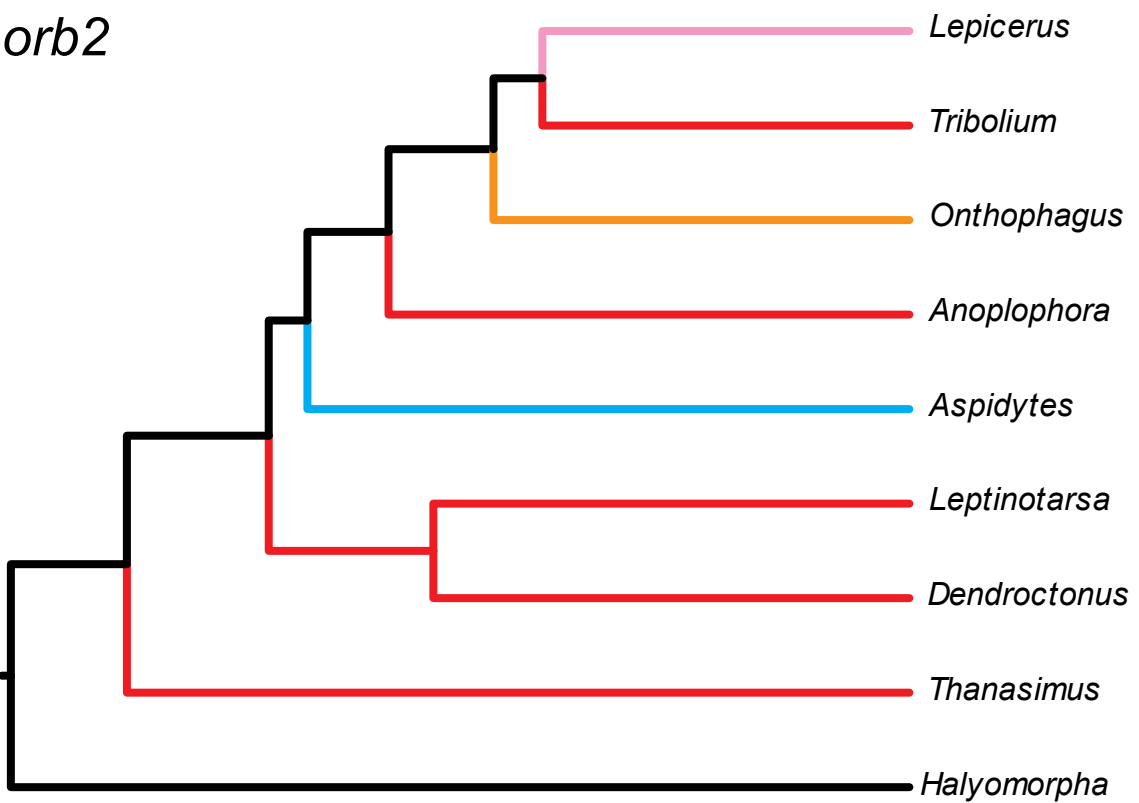
*Npc1a*



*nsr*

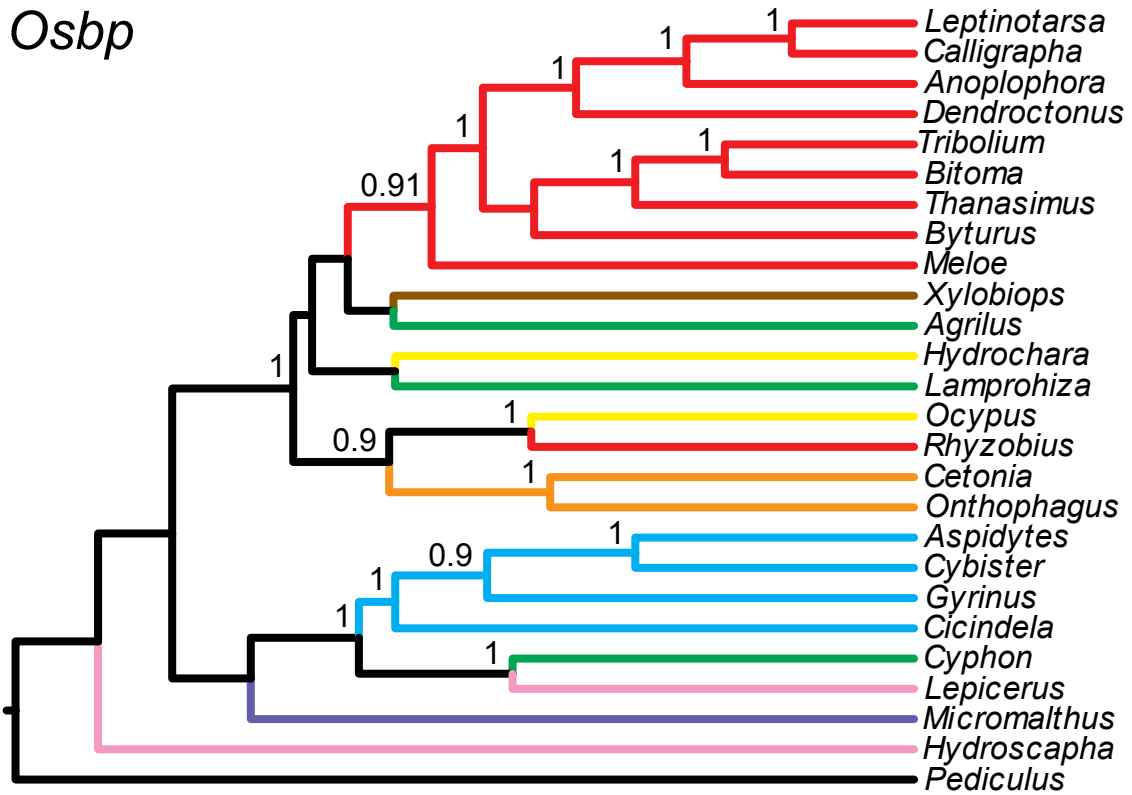


*orb2*

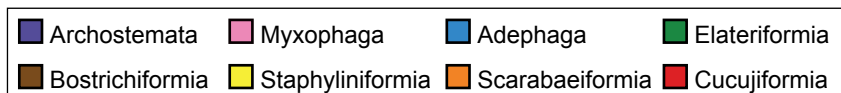
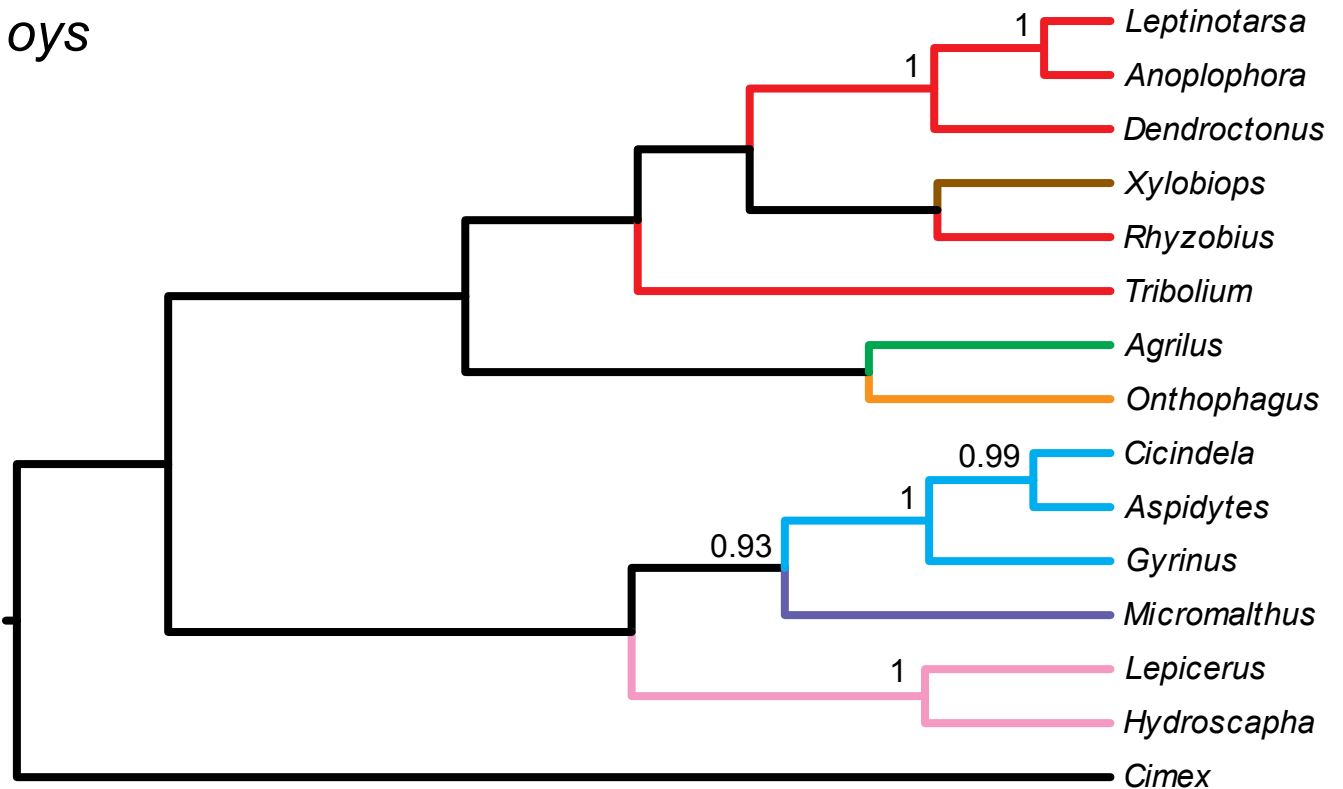




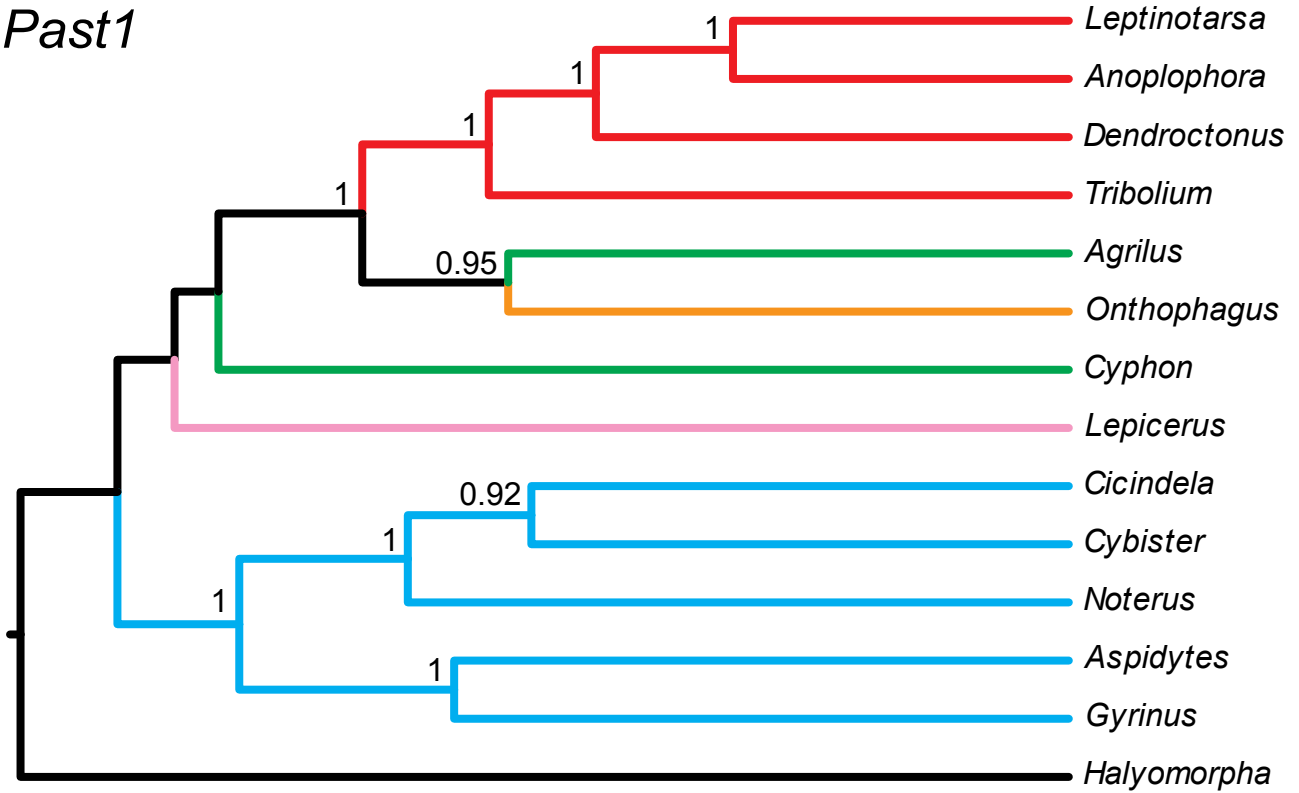
*Osbp*



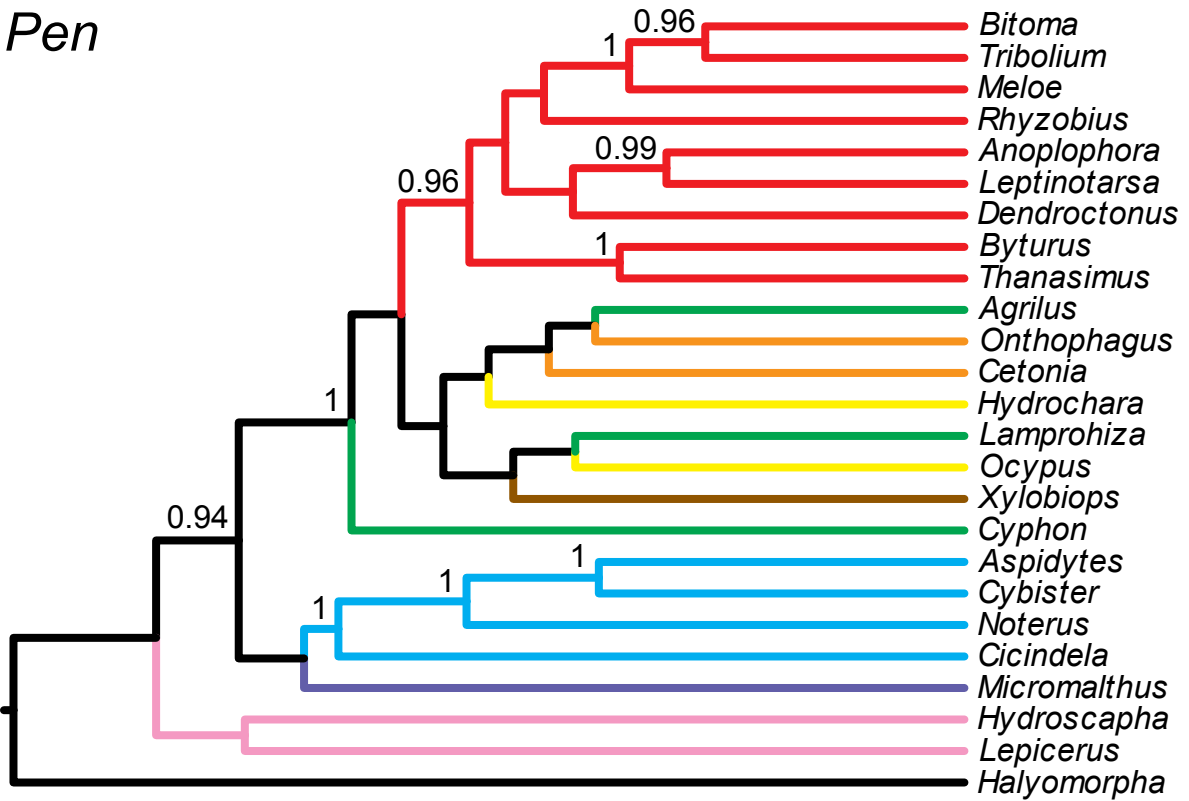
*oys*



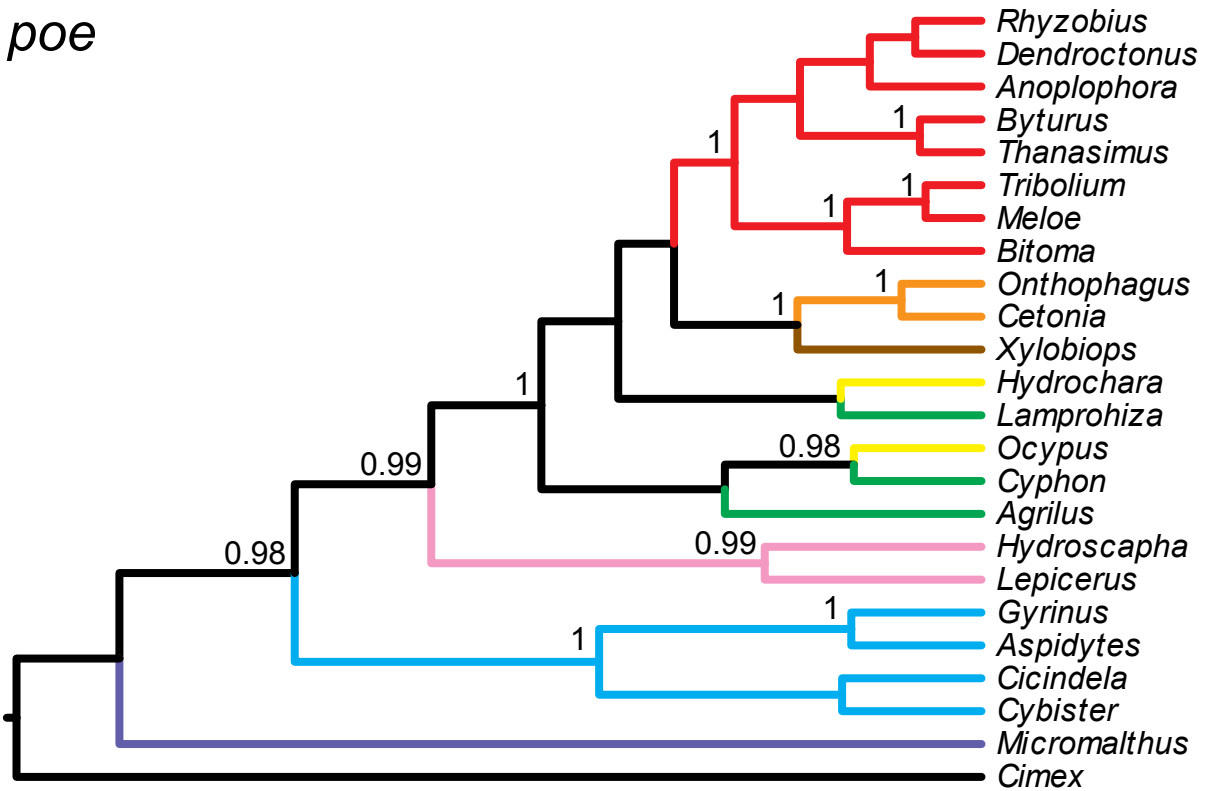
Past1



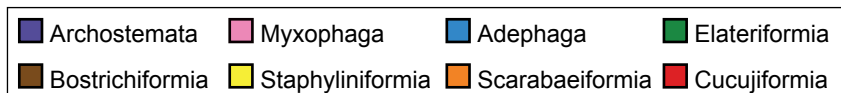
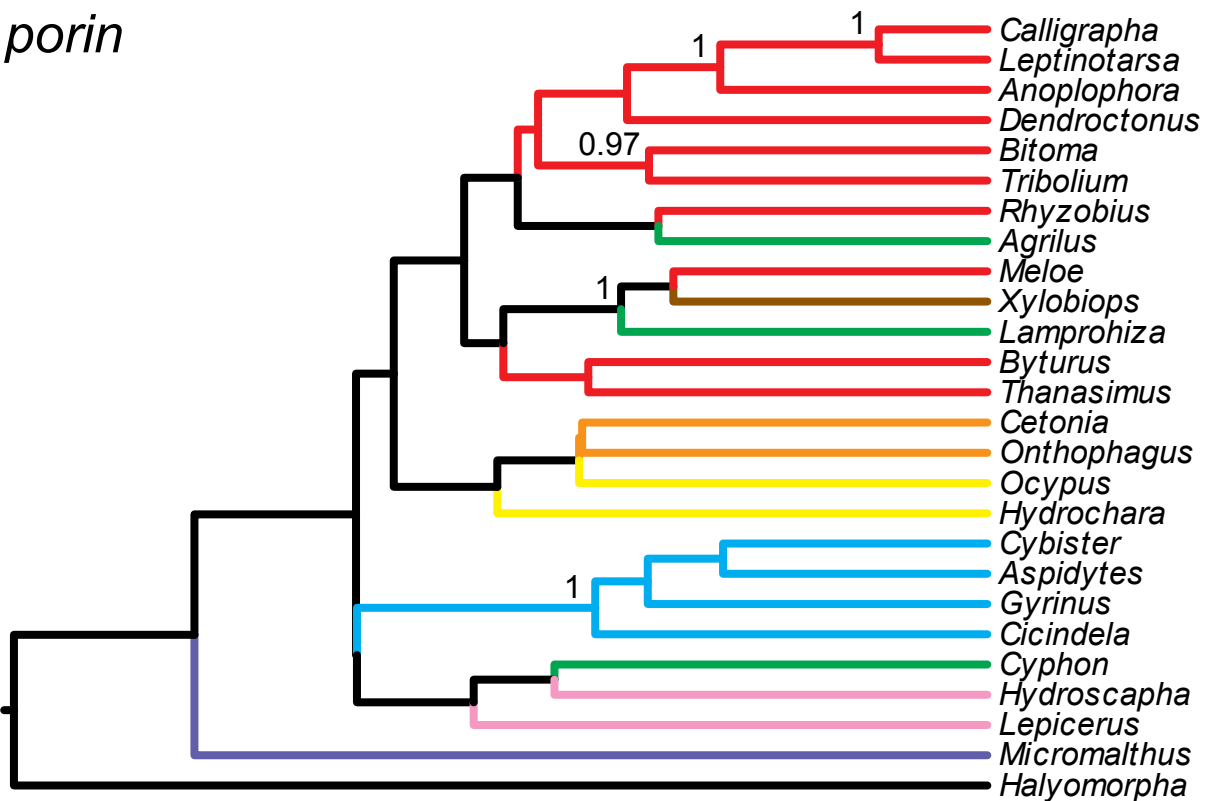
Pen



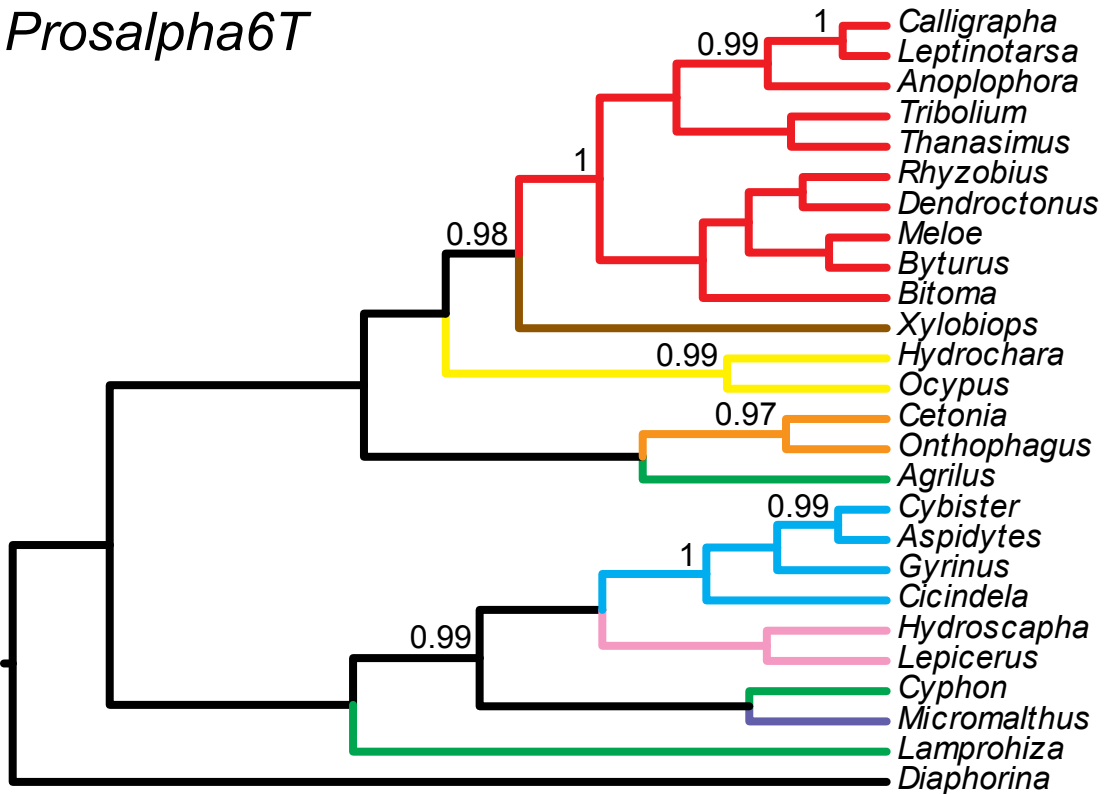
*poe*



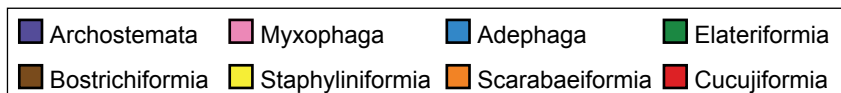
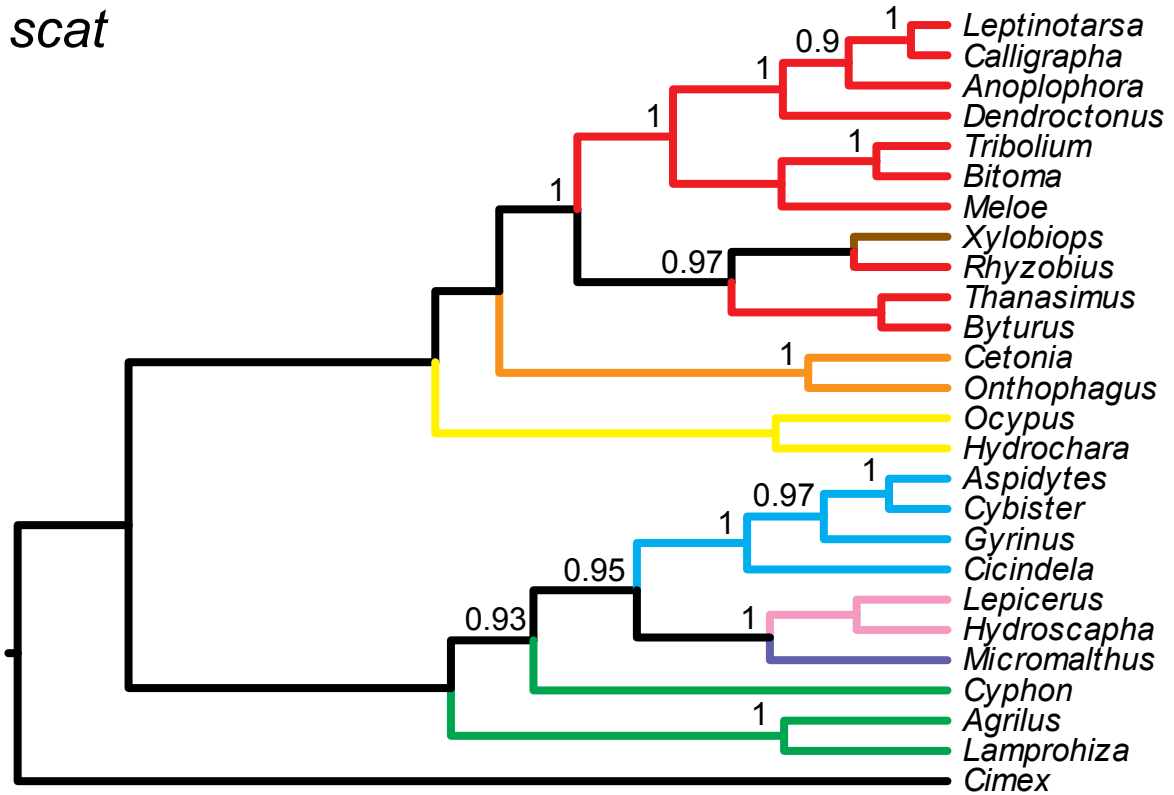
*porin*



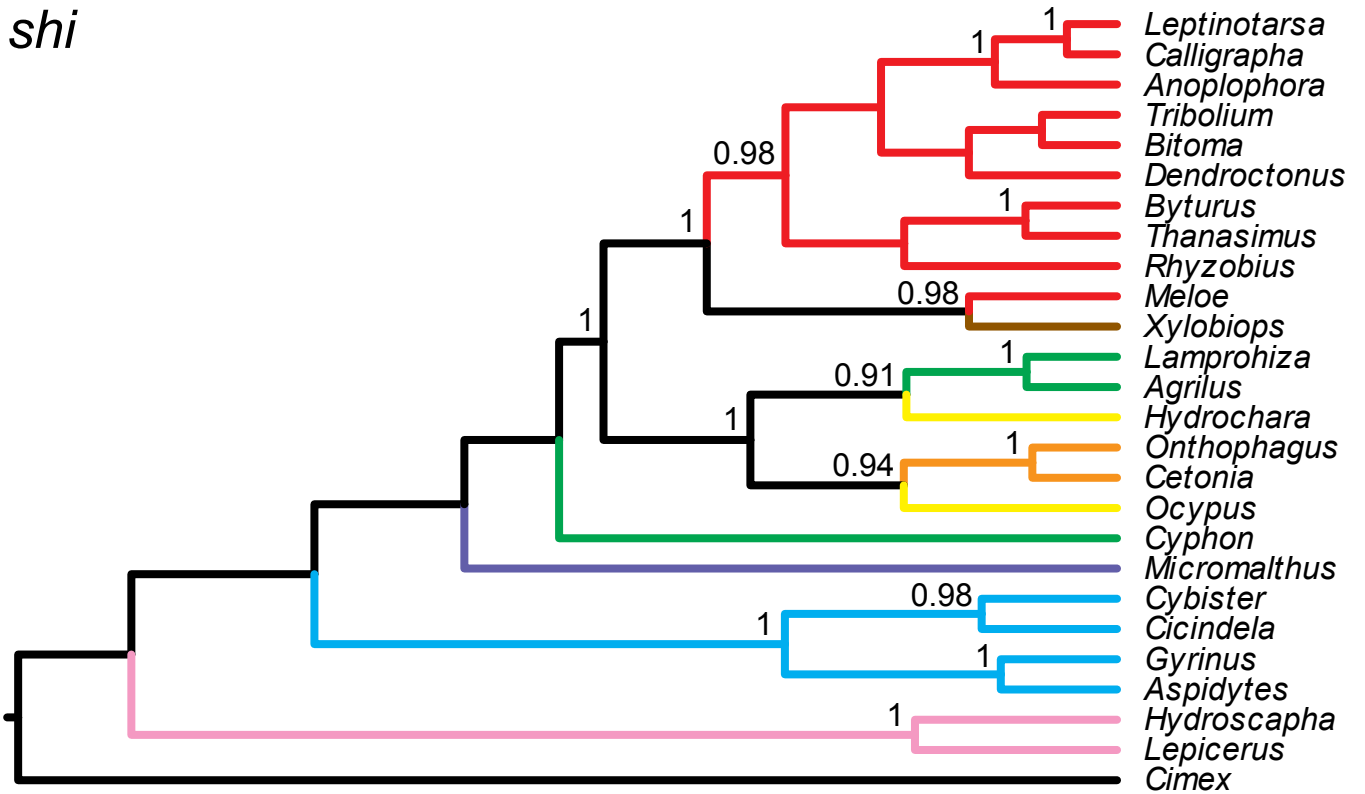
*Prosalpha6T*



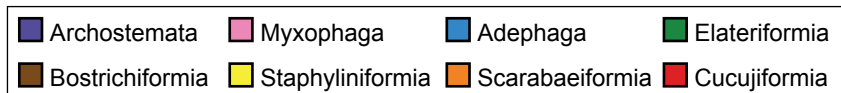
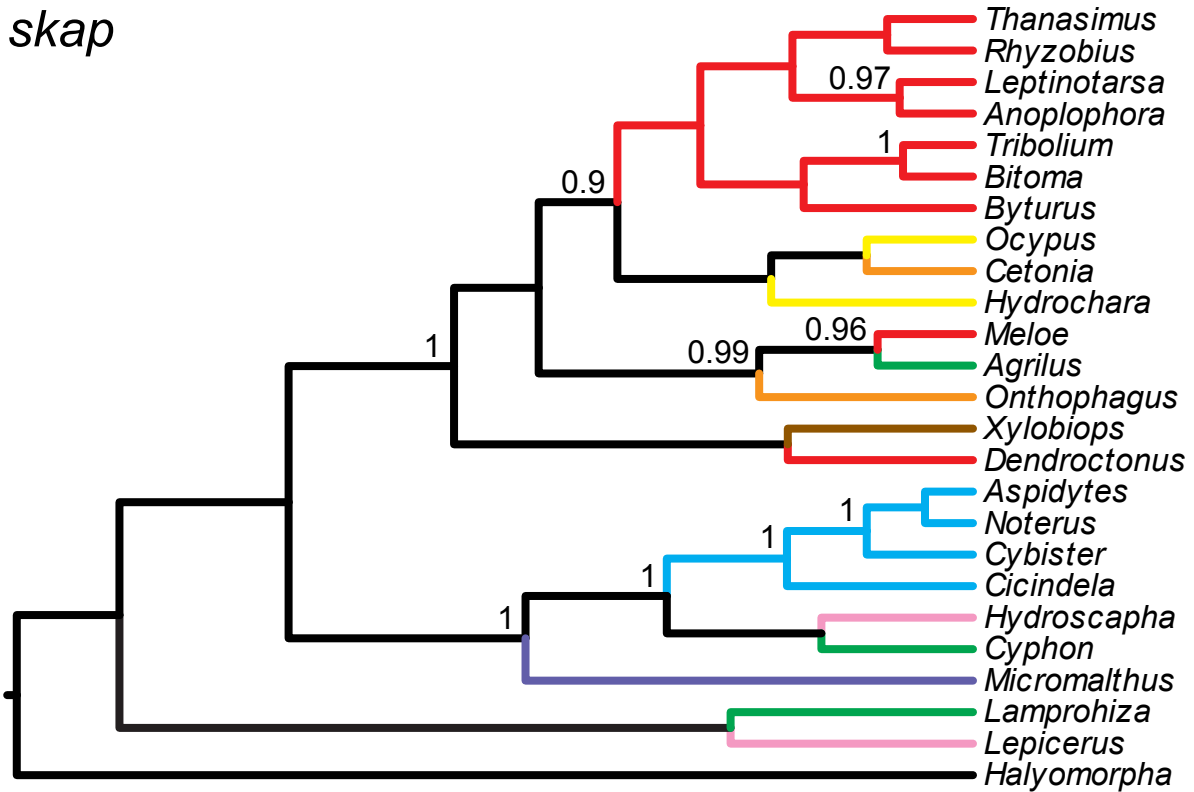
*scat*



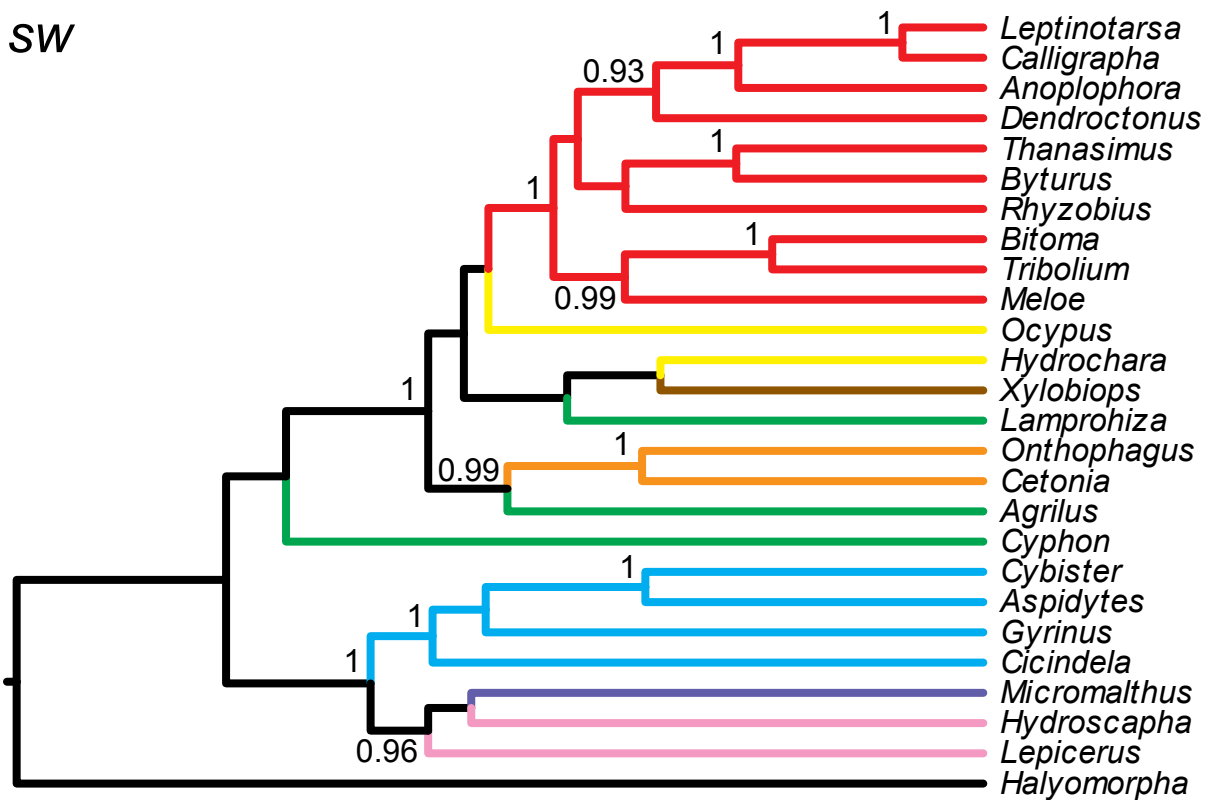
shi



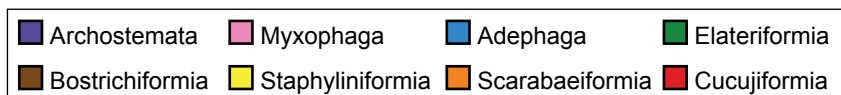
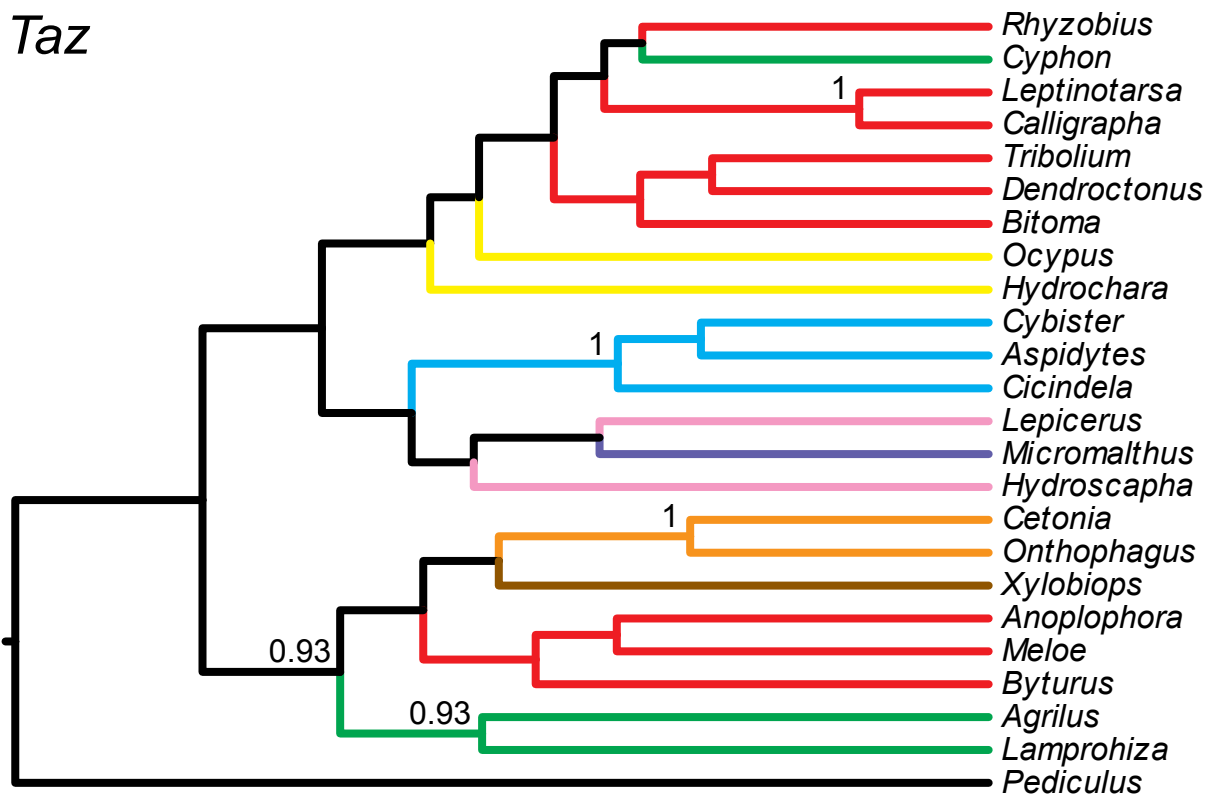
skap



SW



Taz





Vps28

